Crowdsourcing visual perception experiments: A case of contrast threshold (#40944)

First submission

Guidance from your Editor

Please submit by **2 Oct 2019** for the benefit of the authors (and your $200 publishing discount).

Structure and Criteria
Please read the 'Structure and Criteria' page for general guidance.

Custom checks
Make sure you include the custom checks shown below, in your review.

Author notes
Have you read the author notes on the guidance page?

Raw data check
Review the raw data. Download from the location described by the author.

Image check
Check that figures and images have not been inappropriately manipulated.

Privacy reminder: If uploading an annotated PDF, remove identifiable information to remain anonymous.

Files

Download and review all files from the materials page.

1 Figure file(s)

Custom checks

Human participant/human tissue checks

- Have you checked the authors ethical approval statement?
- Does the study meet our article requirements?
- Has identifiable info been removed from all files?
- Were the experiments necessary and ethical?

Structure and Criteria

Structure your review

The review form is divided into 5 sections. Please consider these when composing your review:

1. **BASIC REPORTING**
2. **EXPERIMENTAL DESIGN**
3. **VALIDITY OF THE FINDINGS**
4. General comments
5. Confidential notes to the editor

You can also annotate this PDF and upload it as part of your review

When ready submit online.

Editorial Criteria

Use these criteria points to structure your review. The full detailed editorial criteria is on your guidance page.

BASIC REPORTING

- Clear, unambiguous, professional English language used throughout.
- Intro & background to show context. Literature well referenced & relevant.
- Structure conforms to PeerJ standards, discipline norm, or improved for clarity.
- Figures are relevant, high quality, well labelled & described.
- Raw data supplied (see PeerJ policy).

EXPERIMENTAL DESIGN

- Original primary research within Scope of the journal.
- Research question well defined, relevant & meaningful. It is stated how the research fills an identified knowledge gap.
- Rigorous investigation performed to a high technical & ethical standard.
- Methods described with sufficient detail & information to replicate.

VALIDITY OF THE FINDINGS

- Impact and novelty not assessed. Negative/inconclusive results accepted. *Meaningful* replication encouraged where rationale & benefit to literature is clearly stated.
- All underlying data have been provided; they are robust, statistically sound, & controlled.
- Speculation is welcome, but should be identified as such.
- Conclusions are well stated, linked to original research question & limited to supporting results.

Standout reviewing tips

The best reviewers use these techniques

Tip	*Example*
Support criticisms with evidence from the text or from other sources	*Smith et al (J of Methodology, 2005, V3, pp 123) have shown that the analysis you use in Lines 241-250 is not the most appropriate for this situation. Please explain why you used this method.*
Give specific suggestions on how to improve the manuscript	*Your introduction needs more detail. I suggest that you improve the description at lines 57- 86 to provide more justification for your study (specifically, you should expand upon the knowledge gap being filled).*
Comment on language and grammar issues	*The English language should be improved to ensure that an international audience can clearly understand your text. Some examples where the language could be improved include lines 23, 77, 121, 128 – the current phrasing makes comprehension difficult.*
Organize by importance of the issues, and number your points	*1. Your most important issue* *2. The next most important item* *3. …* *4. The least important points*
Please provide constructive criticism, and avoid personal opinions	*I thank you for providing the raw data, however your supplemental files need more descriptive metadata identifiers to be useful to future readers. Although your results are compelling, the data analysis should be improved in the following ways: AA, BB, CC*
Comment on strengths (as well as weaknesses) of the manuscript	*I commend the authors for their extensive data set, compiled over many years of detailed fieldwork. In addition, the manuscript is clearly written in professional, unambiguous language. If there is a weakness, it is in the statistical analysis (as I have noted above) which should be improved upon before Acceptance.*

Crowdsourcing visual perception experiments: A case of contrast threshold

Kyoshiro Sasaki [1, 2, 3], **Yuki Yamada** [Corresp. 2]

[1] Waseda University, Tokyo, Japan
[2] Kyushu University, Fukuoka, Japan
[3] Japan Society for the Promotion of Science, Tokyo, Japan

Corresponding Author: Yuki Yamada
Email address: yy@artsci.kyushu-u.ac.jp

Crowdsourcing has commonly been used for psychological research but not for studies on sensory perception. This is because in online experiments, one cannot ensure that the rigorous settings required for the experimental environment are replicated. The present study examined the suitability of online experiments of basic visual perception, for example, the contrast threshold. We conducted similar visual experiments in the laboratory and online. Specifically, we employed three experimental conditions. The first was a laboratory experiment, where a small sample of participants (*n* = 24; laboratory condition) completed a task with 10 iterations. The other two conditions were online experiments: participants were either presented with a task without repetition of trials (*n* = 285; online non-repetition condition) or one with 10 iterations (*n* = 166; online repetition condition). The results showed that there was significant equivalence in the contrast thresholds between the laboratory and the online repetition conditions, although a high level of data exclusion was necessary in the online repetition condition. The contrast threshold was significantly higher in the online non-repetition condition than in the laboratory and online repetition conditions. To make crowdsourcing more suitable for investigating the contrast threshold, we should seek the way to reduce data wastage.

Crowdsourcing visual perception experiments: A case of contrast threshold

Kyoshiro Sasaki[1, 2, 3] and Yuki Yamada[2*]

[1] Faculty of Science and Engineering, Waseda University, Tokyo, Japan

[2] Faculty of Arts and Science, Kyushu University, Fukuoka, Japan

[3]Japan Society for the Promotion of Science, Tokyo, Japan

Running head: Crowdsourcing meets Visual Perception

*Correspondence:

Dr. Yuki Yamada

Faculty of Arts and Science, Kyushu University,

744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan.

E-mail: yamadayuk@gmail.com

TEL & FAX: +81-92-802-5837

Abstract

Crowdsourcing has commonly been used for psychological research but not for studies on sensory perception. This is because in online experiments, one cannot ensure that the rigorous settings required for the experimental environment are replicated. The present study examined the suitability of online experiments of basic visual perception, for example, the contrast threshold. We conducted similar visual experiments in the laboratory and online. Specifically, we employed three experimental conditions. The first was a laboratory experiment, where a small sample of participants (n = 24; laboratory condition) completed a task with 10 iterations. The other two conditions were online experiments: participants were either presented with a task without repetition of trials (n = 285; online non-repetition condition) or one with 10 iterations (n = 166; online repetition condition). The results showed that there was significant equivalence in the contrast thresholds between the laboratory and the online repetition conditions, although a high level of data exclusion was necessary in the online repetition condition. The contrast threshold was significantly higher in the online non-repetition condition than in the laboratory and online repetition conditions. To make crowdsourcing more suitable for investigating the contrast threshold, we should seek the way to reduce data wastage.

Keywords: Online experiment, Perception, Vision, Contrast threshold

Introduction

Over the past decade, experiments in psychological research have gone beyond the laboratory. The increasing diversity of research methods and technological advances have increased opportunities for researchers to use resources outside of the laboratory. For example, researchers are using outsourcing services to recruit experimental participants and, often, even commissioning research firms to conduct their surveys and experiments. In addition, "crowdsourcing," based on outstanding technological advances in the digital environment and mobile information devices (for a review, see Stewart, Chandler, & Paolacci, 2017), has become a powerful tool for psychological research.

Crowdsourcing is a method of recruiting large numbers of people and asking them to participate in surveys or experiments via the Internet. Basically, service providers manage an experimenter's task and act as a payment agency. Using crowdsourcing has a number of advantages. The first is its very low cost—participants receive less than $1 for responding to a simple questionnaire or engaging in an easy cognitive task. Second, large (more than 1,000 people) and diverse (in age, gender, and culture) samples can easily employed. The ease of collecting large amounts of diverse data is not only beneficial from the perspective of random sampling but is also helpful for planning experiments and estimating the effect size prior to conducting the experiment (Chrabaszcz, Tidwell, & Dougherty, 2017). Third, it enables researchers to use their time efficiently. With experiments running all hours of the day and night, data from 1,000 people can be obtained within a day or two, although this depends on how many active users are registered with the service. In brief, crowdsourcing enables us to conduct experiments more efficiently.

So far, crowdsourcing has been used for various kinds of experiments and tasks. For example, there are many experimental studies based on self-report questionnaires (e.g., Crangle & Kart, 2015; Garcia, Kappas, Küster, & Schweitzer, 2016; Gottlieb & Lombrozo, 2018; Hurling, Murray, Tomlin et al., 2017; Sasaki, Ihaya, & Yamada, 2017), visual search tasks (de Leeuw & Motz, 2015), reaction time tasks (e.g., Nosek, Banaji, & Greenwald, 2002; Sasaki et al., 2017; Schubert, Murteira, Collins, & Lopes, 2013), keystrokes tasks (Pinet et al., 2017), Stroop tasks (Barnhoorn, Haasnoot, Bocanegra, & van Steenbergen, 2014; Crump, McDonnell, & Gureckis, 2013), the attentional blink task (Barnhoorn et al., 2014; Brown et al., 2014), flanker tasks (Simcox & Fiez, 2014; Zwaan et al., 2018), Simon tasks (Zwaan et al., 2018), lexical decision tasks (Simcox & Fiez, 2014), category learning tasks (Crump et al., 2013), memory tasks (Brown et al., 2014; Zwaan et al., 2018), priming tasks (Zwaan et al., 2018), and decision-making tasks (Berinsky, Huber, & Lenz, 2012; Brown et al., 2014). Moreover, a recent study recruited infants aged five to eight months via crowdsourcing and measured their looking time with webcams (Tran, Cabral, Patel, & Cusack, 2017). These studies suggested that the effect size of the performances in such tasks is comparable to that in laboratory experiments, although the results for the subliminal priming effect (Barnhoorn et al., 2014; Crump et al., 2013) and the cheerleader effect in facial attractiveness (Ojiro et al., 2015) were not consistent between in-laboratory and online experiments. The high reliability of some experimental data was also confirmed (Buhrmester, Kwang, & Gosling, 2011; Crump et al., 2013; de Leeuw & Motz, 2015; Ramsey, Thompson, McKenzie, & Rosenbaum, 2016). Therefore, crowdsourcing can be used for diverse online experiments with publishable reliability.

However, conventional studies on sensory perception are completed in the laboratory. Moreover, it is not unusual that only authors or their laboratory members, who should be well experienced with psychophysical measurements, participate in experiments about sensory perception. Indeed, at this stage, only a small number of studies have tried to run sensory perceptual experiments via crowdsourcing. Previous studies investigated color (Lafer-Sousa, Hermann, & Conway, 2015; Szafir, Stone, & Gleicher, 2014) and randomness (Yamada, 2015) on the web but used one-time color-matching, color word selection, forced choice (same or different), or magnitude estimation tasks. A few studies have measured the point of subjective equality, sensitivity, or thresholds using psychophysical methods in studies on color perception (Ware et al., 2018), volume perception (Pechey et al., 2015), size perception (Brady & Alvarez, 2011) scene perception (Brady, Shafer-Skelton, & Alvarez, 2017), and stimulus visibility (Bang, Shekhar, & Rahnev, 2019). One reason why experiments on sensory perception are rarely conducted online stems from the necessity for rigorous control over the experimental environment. Online experiments depend a great deal on the participant's own computing environment and experimenters cannot control the display settings, visual distance (or the visual field), or the lighting conditions. Thus far, online experiments seem inadequate for experimental studies that focus on the visual functions of spatial and temporal resolutions such as contrast threshold. For example, previous studies examined the issue of the temporal aspect of stimulus presentation and found that stimuli were systematically presented for 20 ms longer than the programmed durations (de Leeuw & Motz, 2015; Reimers & Stewart, 2014).

In the present study, we focused on measuring low-level visual perception via online experiments. We examined the contrast threshold in vision via online crowdsourcing and the laboratory. The contrast threshold is a non-temporal visual capacity that is most susceptible to the influence of the display condition in measurement. This is because its measurement needs strict linearization of the outputs of the display with gamma correction; however, most displays of home PCs are not linearized. We believed that a comparison between web and lab measurements of visual contrast thresholds would provide tangible evidence of what online experiments can and cannot test regarding non-temporal aspects of stimulus presentation. If the non-linearity of monitor displays were a negligible random effect, the contrast threshold online and in the laboratory would be similar. Moreover, the present study used two types of iteration for online experiments, that is, the repetition and non-repetition conditions. In the former, participants were presented with each trial 10 times per stimulus condition but in the latter condition, each trial was presented only once. In online experiments, boredom in participants substantially decreases data quality (Chandler, Mueller, & Paolacci, 2014) and many repetitions are likely to induce boredom. If we could control for measurement errors or individual differences by increasing the sample size, a single trial for a stimulus condition would be enough to lead to an appropriate conclusion, even in online experiments, without data deterioration. For this reason, the sample size of participants in the non-repetitive condition was about 10 times that of the repetitive condition.

Methods

Participants. We used G*power to determine the sample sizes needed for the repetition condition (α = .05, 1-β = .80). In the laboratory condition, we used a moderate effect size (*f* = .25) in the calculation of the required sample size. The required and maximum sample size was 24. In the online repetition condition, we used a small effect size (*f* = .10) in the calculation of the required sample size, because of the potential for noise in the data of online experiments. The required sample size was 138. Considering potential satisficers (Chandler et al., 2014; Oppenheimer, Meyvis, & Davidenko, 2009), 200 people was set as the maximum sample size and participants were recruited through a crowdsourcing service (Yahoo! Crowdsourcing: http://crowdsourcing.yahoo.co.jp/). We considered that the required sample size in the online non-repetition condition should be at the least 10 times number of that in the laboratory condition (240 people) according to the differences in the number of repetitions. Similarly, in the online repetition condition, we recruited 300 people as the maximum sample size to account for the potential influence of satisficers. The participants in the laboratory conditions undertook several experiments, including the present experiment, for 3 hours and received 4000 JPY. The participants in the online repetition and non-repetition conditions got 50 and 20 T-points (1 T-point = 1 JPY)[1]. Participants were not made aware of the purpose of the study. The experiment was conducted according to the principles laid down in the Helsinki Declaration. The protocol was approved by the ethics committees of Waseda University (approval number: 2015-033) and Kyushu University (approval number: 2016-017). We obtained written informed consent from all the participants in the laboratory condition. On the other hand, it was difficult to obtain written informed consent in the online conditions. Thus, according to the protocol (approval number: 2016-017), we explained the details of the online experiments by the instruction beforehand, and then asked the participants to take part in the experiments only when they agreed to the instruction.

Apparatus. In the laboratory condition, stimuli were presented on a 23.5-inch LCD display (FG2421; EIZO, Japan). The resolution of the display was 1920 × 1080 pixels and the refresh rate was 100 Hz. We performed gamma correction for the luminance emitted from the monitor. The presentation of stimuli and the collection of data were computer-controlled (Mac mini, Apple, USA). We used MATLAB with the Psychtoolbox extension (Brainard, 1997; Pelli, 1997) to generate the stimuli. The observer's visual field was fixed using a chin-head rest at a viewing distance of 57 cm. The size information at the visual angle described for the laboratory condition was based on this viewing distance. In the online conditions, the experiment was conducted on a web browser with a JavaScript application (jsPsych; de

Leeuw, 2015).

Stimuli and Procedure. Stimuli consisted of a fixation circle (diameter was 0.24 degrees) and Gabor patches, the diameter of which was 42 pixels (2 degrees in the laboratory conditions). The SD of a gaussian function was 6 pixels (0.29 degrees). There were four spatial frequencies of the carrier: 0.02, 0.05, 0.09, and 0.38 cycles per pixel (cpp; 0.5, 1, 2, and 8 cycles per degree [cpd] in the laboratory conditions). We set seven contrast levels (the Michelson contrast) varying across the spatial frequencies. The contrasts in the 0.02 cpp (0.5 cpd) trials were 3, 8, 13, 18, 23, 28, and 33%. The contrasts in the 0.05 and 0.09 cpp (1 and 2 cpd) trials were 1, 6, 11, 16, 21, 26, and 31%. The contrasts in the 0.38 cpp (8 cpd) trials were 5, 10, 15, 20, 25, 30, and 35%. The Gabor patches were tilted 45° clockwise or counterclockwise. We took screenshots of the stimuli on the monitor at the laboratory and used them for the online conditions.

In the laboratory condition, the experiment was conducted in a dark room. The participants initiated each trial by pressing the space key. The fixation circle was presented for 500 ms. After the fixation circle disappeared, the Gabor patch was presented for 50 ms, Then, we presented a blank screen for 300 ms, followed by the prompt: “In which direction was the stimulus tilted?” The method of constant stimuli was used. The participants were asked whether the stimulus was tilted clockwise or counterclockwise. Participants responded without time limits or feedback. Each of the spatial frequency conditions was conducted in a separate session; thus, the experiment consisted of 4 sessions. The session order was randomized across participants. In each session, trials were conducted for 7 contrasts in 2 orientations. In the repetition condition, each combination of contrast and orientation was presented 10 times per session. Thus, participants in the repetition condition completed 560 trials in total, whilst those in the non-repetition condition completed 56. The order of the trials was also randomized across participants. Before the first session, we conducted a practice session, where the participants completed 4 trials. The spatial frequency of the practice session was identical to that of the first session, and the contrast was 100%. Both of the orientations appeared twice. The trial order of each session was randomized across the participants.

In the online conditions, the procedures were identical to that of the laboratory conditions except for the added insertion of attention check questions (ACQs). This is because online participants are often distracted (Chandler et al., 2014) or are satisficers (Oppenheimer et al., 2009), and previous studies have found that ACQs can reduce low-quality responses (Aust, Diedenhofen, Ullrich, & Musch, 2012; Oppenheimer et al., 2009). The ACQs were easy calculations based on the four basic arithmetic operations (e.g., 20 + 15 = ?) and participants selected the correct answer from five options, ACQs appeared halfway through the total number of trials in each session.

Data Analysis. We excluded participants who gave incorrect answers to one or more ACQs. In the laboratory and online repetition conditions, we calculated the contrast threshold of each spatial frequency for each participant, at which the proportion of "correct" responses was 0.82 (Cameron, Tai, & Carrasco, 2002; Lee, Baek, Lu, & Mather, 2014), using a probit analysis (i.e., fitting a cumulative Gaussian function to the proportion of "correct" responses as a function of the contrast level). We used the "glm" function in R (3. 4. 4). The probit analysis provided the means and standard deviations (SDs) of the distributions. Then, we calculated the contrast thresholds using the means, SDs, and the "qnorm" function in R. We excluded participant data when β calculated by the probit analysis was a negative value. This was because the negative value indicated a reduction in correct responses as the contrast level increased, and in this case, the participants could not perform the task precisely. As a result, we could not calculate the thresholds. We also excluded the data of participants whose contrast thresholds were under 0 or over 100% because the contrast threshold should be within this range. In the online non-repetition conditions, we used the pooled data from all the participants and calculated the contrast threshold for each spatial frequency by the same procedures of the repetition conditions.

First, to confirm whether the contrast threshold depended on the spatial frequency, we conducted a one-way analysis of variance (ANOVA) on the contrast thresholds with spatial frequency as a within-participant factor, for the laboratory and online repetition conditions. We set the alpha level as .05 and calculated η_p^2. When the main effects were significant, we conducted multiple comparison tests using Holm's method (Holm, 1979). We had to conduct the *t*-tests six times. Therefore, we increased α from .008 to .05 based on Holm's correction (Holm, 1979).

Moreover, our purpose was to examine whether the contrast thresholds were different or equivalent between experimental environments in each spatial frequency. Thus, we conducted two-tailed Welch's *t*-tests for the contrast thresholds for each spatial frequency. After the *t*-tests, we conducted equivalence tests for the pairs in which the contrast thresholds were not significantly different. For the equivalence tests, we used the TOSTER package in R (Lakens, Scheel, & Isager, 2018) and set Cohen's *d* to 0.5. We compared the contrast threshold of the laboratory condition and the online repetition and non-repetition conditions, and thus we had to conduct *t*-tests and equivalence test three times at most. Therefore, we set α from .017 to .05 based on Holm's correction (Holm, 1979).

Results

The results are shown in Figure 1. We collected the data from 24 people in the laboratory condition. In the online repetition condition, of the 200 people recruited, only 80 participated. As this number did not reach the required sample size, we recruited another 200 people and subsequently collected data from 166 participants. For the online non-repetition condition, of the 300 people recruited, only 156 participated. Therefore, we recruited another 250 people and subsequently collected data from 285 participants. We excluded the data of 2, 84, and 19 participants in the laboratory, online repetition, and online non-repetition conditions, respectively, based on the rules detailed in the *Data Analysis* section.

Online repetition and Laboratory Conditions. The results of the ANOVA on the contrast thresholds in the laboratory condition revealed that the main effect was significant, $F(3, 63) = 7.63$, $p < .001$, $\eta_p^2 = .27$. The multiple comparison tests showed that the threshold was significantly higher in the 0.5 cpd trials than in the 1 and 2 cpd trials, $ts(21) > 6.25$, $ps < .001$, Cohen's $dzs > 1.33$. Moreover, the threshold was significantly higher in the 4 cpd trials than in the 2 cpd trials, $t(21) = 2.88$, $p = .009$, Cohen's $dz = 0.61$. The results of the ANOVA on the contrast thresholds in the online repetition condition revealed that the main effect was significant, $F(3, 243) = 26.23$, $p < .001$, $\eta_p^2 = .24$. The multiple comparison tests showed that the threshold was significantly higher in the 4 cpd trials than in the 1 and 2 cpd trials, $ts(81) > 6.77$, $ps < .001$, Cohen's $dzs > 0.74$. The threshold was also significantly higher in the 0.5 cpd trials than in the 1 and 2 cpd trials, $ts(81) > 4.98$, $ps < .001$, Cohen's $dzs > 0.64$.

Online non-repetition, Online repetition, and Laboratory Conditions. For the 0.5 cpd trials, the threshold was significantly higher in the online non-repetition condition than in the online repetition, $t(332.97) = 6.14$, $p < .001$, Cohen's $d = 0.51$, and laboratory, $t(159.41) = 5.95$, $p < .001$, Cohen's $d = 0.45$, conditions, while there was no significant difference between the online repetition and laboratory conditions, $t(68.92) = 0.31$, $p = .76$, Cohen's $d = 0.05$. The equivalence test showed significant equivalence between the online repetition and laboratory conditions, $t(68.92) = 2.26$, $p = .013$.

For the 1 cpd trials, the threshold was significantly higher in the online non-repetition condition than in the online repetition, $t(314.58) = 7.54$, $p < .001$, Cohen's $d = 0.55$, and laboratory, $t(285.95) = 7.43$, $p < .001$, Cohen's $d = 0.71$, conditions, while there was no significant difference between the online repetition and laboratory conditions, $t(82.43) = 0.56$, $p = .580$, Cohen's $d = 0.09$. The equivalence test showed significant equivalence between the online repetition and laboratory conditions, $t(82.43) = 2.13$, $p = .018$.

For the 2 cpd trials, the threshold was significantly higher in the online non-repetition condition than in the online repetition, $t(319.24) = 7.06$, $p < .001$, Cohen's $d = 0.52$, and laboratory, $t(268.92) = 7.11$, $p < .001$, Cohen's $d = 0.72$, conditions, while there was no significant difference between the online repetition and laboratory conditions, $t(57.31) = 0.33$, $p = .742$, Cohen's $d = 0.06$. The equivalence test showed significant equivalence between the online repetition and laboratory conditions, $t(57.31) = 2.12$, $p = .019$.

For the 4 cpd trials, the threshold was significantly higher in the online non-repetition condition than in the online repetition, $t(344.97) = 6.23$, $p < .001$, Cohen's $d = 0.50$, and laboratory, $t(56.41) = 5.06$, $p < .001$, Cohen's $d = 0.51$, conditions, while there was no significant difference between the online repetition and laboratory conditions, $t(31.40) = .564$, $p = .577$, Cohen's $d = 0.14$. The equivalence test showed that the equivalence between the online repetition and laboratory conditions was marginally significant, $t(31.40) = 1.48$, $p = .075$.

Discussion

In the present study, we examined whether the contrast threshold was properly measured in an online experiment with two conditions: a condition with repetition of trials and one without repetition. The results showed that there were equivalences in the contrast thresholds of the online repetition and laboratory conditions. On the other hand, at high spatial frequencies, the contrast threshold in the online non-repetition condition was higher than that in the online repetition and laboratory conditions. Thus, online experiments seem to be able to measure the contrast threshold as adequately as laboratory experiments, provided there are enough repetitions; it is difficult to measure the contrast thresholds without repetitions. Additionally, we have to discuss several points to improve the suitability of crowdsourcing for online perceptual experiments.

Is crowdsourcing suitable for investigating the contrast threshold? The present study excluded 51% of the data in the online repetition condition. These exclusions mainly stemmed from the fact that the correct response decreased as the contrast level increased or the thresholds were under zero. That is, in the online repetition condition, it was rare to precisely calculate the contrast threshold. One possibility is that the experimental environment of 49% of the participants in the online repetition condition might be similar to that of the laboratory condition. As a result, we were able to calculate the thresholds of these participants and we found significant equivalences between the laboratory and online repetition conditions. On the other hand, the contrast thresholds were much higher in the online non-repetition condition. Although it is difficult to interpret this result, the results of the online repetition condition may inform our understanding. A large amount of the data was excluded in the online repetition condition. Based on this, we can expect the data obtained via online experiments to be noisy. This noisy data should be included in the online non-repetition condition and mediate in the results of the online non-repetition condition. Considering the exclusion of data in the online repetition condition and the results of the online non-repetition condition, we cannot conclude that online experiments are adequate for measuring the contrast threshold; in fact, it will be difficult to measure the contrast threshold via crowdsourcing unless we are able measure and calibrate the lighting conditions of each online participant via camera.

There may be some solutions for improving the situation of online measurements of the contrast threshold. One solution would be to control the experimental environments of each participants in the online experiments and make them similar to that of a laboratory experiment. A previous study proposed beneficial tips for controlling the size of stimuli, distance from the monitor, sound volume, and brightness (Woods, Velasco, Levitan, Wan, & Spence, 2015). Woods et al. also provided a possible way to adjust color, which seems to be difficult to control across online participants, using the psychophysical method (To, Woods, Goldstein, & Peli, 2013). At this time, these methods seem to require much effort from the participants and thus might not be effective. The growth of ways to control experimental environments should lead to a reduction in low-quality data, leading to a decreased in the exclusion of data.

The other way is related to participant negligence. In the online experiment, it was difficult for participants to maintain their motivation while performing tasks, because they were not prepared to participate in psychological experiments and did not face experimenters. In such situations, participants often do not devote enough effort to the tasks and, hence, cursory responses increase (Satisficing: e.g., Berinsky, Margolis, & Sances, 2016; Maniaci & Rogge, 2014; Miura & Kobayashi, 2016; Oppenheimer et al., 2009). ACQs, which we set during the online condition sessions, are beneficial for protecting the quality of the data from satisficing. It is easy for participants to correctly answer ACQs when they perform the tasks carefully. Generally, it is important to exclude the data of those who wrongly answer ACQs because of inattention and/or cursory responses in order to improve the quality of the data.

However, in the present study, the data exclusion due to incorrect ACQ responses accounted for 6% of the total data in each of the online conditions. Thus, the ACQ might not be working as intended in the present study. The type of ACQ was extremely different from that of the main task (i.e., judging the orientation of the Gabor patch). Given this, perhaps we should improve the ACQ so that participants are not easily caught out, or we should use another method; an instructional manipulation check (IMC) is also helpful for detecting satisficers (Oppenheimer et al., 2009). An IMC checks whether the participants carefully read the instructions for the tasks. Specifically, they can incorporate the instruction not to answer the questions into some methods commonly used in psychological research (e.g., Likert scales); thus, if the participants do not carefully read the instructions, they mistakenly answer the questions. The data of such participants should be excluded because they improperly dealt with the tasks. Additionally, a recent study showed that alerting satisficers to their inattentiveness by a repeated IMC was helpful for improving their information processing (Miura & Kobayashi, 2016). In general, ACQs and IMCs are valid tools for the detection and exclusion of data from satisficers. However, it is difficult to prevent satisficers from participating in experiments. To keep from losing data owing to satisficers, blacklisting them might be more effective in the long term.

In addition, there may be other ways to maintain the quality of psychophysical online data. One possibility is developing a platform designed for scientific research. Crowdsourcing services such as Yahoo! Crowdsourcing and Amazon Mechanical Turk have some advantages for conducting psychological research. However, they were not developed as research tools and have some inconveniences as well. Recently, a platform for scientific research was designed (TurkPrime: Litman, Robinson, & Abberbock, 2017) and integrated with Amazon Mechanical Turk. Some helpful systems for improving the quality of online data were provided, including: excluding participants based on previous participation, communicating with participants, monitoring dropout and engagement rates. Elevating these platforms should be helpful for improving the quality of the data in the online experiments.

Crowdsourcing enables us to obtain large amounts of data from various people. This is advantageous for examining individual differences in perceptual and cognitive processing. In classic laboratory experiments, most participants are university or graduate students. It is also difficult to collect large amounts of data in classic laboratory experiments. Thus, the personal traits of the participants do not vary enough to examine the relation between individual differences in perceptual and cognitive processing. Crowdsourcing, however, allows researchers to recruit participants from around the world, and hence we can collect mass data from participants with various personality traits. Indeed, we and others have already shown the relationship between individual differences in personality traits and perceived eeriness using crowdsourcing (Chaya et al., 2015; Sasaki et al., 2017). Moreover, we previously conducted a perceptual study indicating the individual differences in the perception of pattern randomness (Yamada, 2015). If the environment in the online experiments is improved and crowdsourcing becomes suitable, to some extent, for investigating visual perception, online experiments will be helpful for addressing individual differences in visual perception.

Footnote

[1]We paid 82 and 37 JPY to Yahoo! Crowdsourcing for every participant in the online repetition and non-repetition conditions, respectively. Yahoo! Crowdsourcing received the difference between the participant's reward and our payment as a profit margin.

References

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2012). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535. doi: 10.3758/s13428-012-0265-2

Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency. *Journal of Experimental Psychology: General*, *148*, 437–452. doi: 10.1037/xge0000511

Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2014). QRTEngine: An easy solution for running online reaction time experiments using Qualtrics. *Behavior Research Methods*, 1–12. doi: 10.3758/s13428-014-0530-7

Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368. doi: 10.1093/pan/mpr057

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers? *Journal of Experimental Social Psychology*, *66*, 20–28. doi: 10.1016/j.jesp.2015.09.010

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*, 384–392.

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*, 1160–1176.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436. doi: 10.1163/156856897X00357

Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., & Dolan, R. J. (2014). Crowdsourcing for cognitive science – The utility of smartphones. *PLoS ONE, 9* (7), e100662. doi: 10.1371/journal.pone.0100662

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* (1), 3–5. doi: 10.1177/1745691610393980

Cameron, E. L., Tai, J. C., & Carrasco, M. (2002). Covert attention affects the psychometric function of contrast sensitivity. *Vision Research*, *42*, 949–967.

Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk workers: consequences and solutions for behavioral researchers. *Behavior Research Methods*, *46* (1), 112–130. doi: 10.3758/s13428-013-0365-7

Chaya, K., Xue, Y., Uto, Y., Yao, Q., & Yamada, Y. (2016). Fear of eyes: Triadic relation among social anxiety, trypophobia, and discomfort for eye cluster. *PeerJ*, *4*:e1942. doi: 10.7717/peerj.1942

Chrabaszcz, J. S., Tidwell, J. W., & Dougherty, M. R. (2017). Crowdsourcing prior information to improve study design and data analysis. *PLOS ONE*, *12*: e0188246. doi: 10.1371/journal.pone.0188246

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Crangle, C. E., & Kart, J. B. (2015). A questions-based investigation of consumer mental-health information. *PeerJ*, *3*:e867. doi: 10.7717/peerj.867

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* (3), 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.

Garcia, D., Kappas, A., Küster, D., & Schweitzer, F. (2016). The dynamics of emotions in online interaction. *Royal Society Open Science*, *3*:8. doi: 10.1098/rsos.160059

Gottlieb, S., & Lombrozo, T. (2018). Can Science Explain the Human Mind? Intuitive Judgments About the Limits of Science. *Psychological Science*, *29*, 121–130. doi: 10.1177/0956797617722609

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*, 65–70. doi: 10.2307/4615733

Hurling, R., Murray, P., Tomlin, C., Warner, A., Wilkinson, J., York, G., Linley, P. A., Dovey, H., Hogan, R. A., Maltby, J., & So, T. T. (2017). Short Tips Delivered "in the Moment" Can Boost Positive Emotion. *International Journal of Psychological Studies*, *9*, 88–106. doi: 10.5539/ijps.v9n1p88

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by "the dress" photograph. *Current Biology*, *25*, R1–R2. doi: 10.1016/j.cub.2015.04.053

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for *Psychological* Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*, 259–269.

Lee, T. H., Baek, J., Lu, Z. L., & Mather, M. (2014). How arousal modulates the visual contrast sensitivity function. *Emotion*, *14*, 978–984.

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*, 433–442. doi: 10.3758/s13428-016-0727-z

Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality*, *48*, 61–83. doi: 10.1016/j.jrp.2013.09.008

Miura, A., & Kobayashi, T. (2016). Survey satisficing inflates stereotypical responses in online experiment: The case of immigration study. *Frontiers in Psychology*, *7*:1563. doi: 10.3389/fpsyg.2016.01563

Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory*, *6* (1), 101–115. doi: 10.1037/1089-2699.6.1.101

Ojiro, Y., Gobara, A., Nam, G., Sasaki, K., Kishimoto, R., Yamada, Y., & Miura, K. (2015). Two replications of "Hierarchical encoding makes individuals in a group seem more attractive (2014; Experiment 4)". *The Quantitative Methods for Psychology*, *11*, r8-r11. doi: 10.20982/tqmp.11.2.r008

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology*, *45* (4), 867–872. doi: 10.1016/j.jesp.2009.03.009

Pechey, R., Attwood, A. S., Couturier, D. L., Munafò, M. R., Scott-Samuel, N. E., Woods, A., & Marteau, T. M. (2015). Does glass size and shape influence judgements of the volume of wine? *PLOS ONE*, *10*: e0144536.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442. doi: 10.1163/156856897X00366

Pinet, S., Zielinski, C., Mathot, S., Dufau, S., Alario, F. X., & Longcamp, M. (2017). Measuring sequences of keystrokes with jsPsych: Reliability of response times and inter-keystroke intervals. *Behavior Research Methods*, *49*, 1163–1176.

Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of commodity computers: new hardware, new problems? *Behavior Research Methods*, *41* (3), 598–614. doi: 10.3758/s13428-016-0776-3

Ramsey, S. R., Thompson, K. L., McKenzie, M., & Rosenbaum, A. (2016). Psychological research in the internet age: The quality of web-based data. *Computers in Human Behavior*, *58*, 354–360.

Reimers, S., & Stewart, N. (2014). Presentation and response timing accuracy in Adobe Flash and HTML5/JavaScript Web experiments. *Behavior Research Methods*, 1–19. doi: 10.1016/j.chb.2015.12.049

Sasaki, K., Ihaya, K., & Yamada, Y. (2017). Avoidance of novelty contributes to the uncanny valley. *Frontiers in Psychology*, *8:* 1792. doi: 10.3389/fpsyg.2017.01792

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* (6), 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

Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing Samples in Cognitive Science. *Trends in Cognitive Sciences*, *21*, 736–748. doi: 10.1016/j.tics.2017.06.007

Szafir, D. A., Stone, M., & Gleicher, M. (2014). Adapting color difference for design. *Proceedings of Color and Imaging Conference*, *2014*, 228–233.

To, L., Woods, R. L., Goldstein, R. B., & Peli, E. (2013). Psychophysical contrast calibration. *Vision Research*, *90*, 15–24. doi: 10.1016/j.visres.2013.04.011

Tran, M., Cabral, L., Patel, R., & Cusack, R. (2017). Online recruitment and testing of infants with Mechanical Turk. *Journal of Experimental Child Psychology*, *156*, 168–178. doi: 10.1016/j.jecp.2016.12.003

Turpin, A., Lawson, D. J., & McKendrick, A. M. (2014). PsyPad: a platform for visual psychophysics on the iPad. *Journal of Vision*, *14* (3):16, 1–7. doi: 10.1167/14.3.16

Ware, C., Turton, T. L., Bujack, R., Samsel, F., Shrivastava, P., & Rogers, D. H. (2018). Measuring and Modeling the Feature Detection Threshold Functions of Colormaps. *IEEE transactions on visualization and computer graphics*. doi: 10.1109/TVCG.2018.2855742

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

Yamada, Y. (2015). Gender and age differences in visual perception of pattern randomness. *Science Postprint*, *1* (2): e00041. doi: 10.14340/spp.2015.01A0002

Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, 25, 1968–1972. doi: 10.3758/s13423-017-1348-y

Acknowledgments

We would like to thank Dr. Daiichiro Kuroki for developing the program of the online experiment. This research was supported by JSPS KAKENHI #17J05236 to K.S. and #15H05709, #16H03079, #16H01866, #17H00875, #18H04199, and #18K12015 to Y.Y.

Competing interests

The authors declare no competing interests.

Author contributions

Contributed to conception and design: KS, YY

Contributed to acquisition of data: KS

Contributed to analysis and interpretation of data: KS

Drafted and/or revised the article: KS, YY

Approved the submitted version for publication: KS, YY

Data accessibility statement

The dataset is shown in https://figshare.com/s/c067967c1cfd3238244b

Figure legend

Figure 1. The results of the laboratory and online experiments. Error bars denote standard deviations.

Figure 1

The results of the experiment

The results of the laboratory and online experiments. Error bars denote standard deviations