Crowdsourcing visual perception experiments: A case of contrast threshold

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

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. A reason is that 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 on basic visual perception, particularly the contrast threshold. We conducted similar visual experiments in the laboratory and online, employing 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 significant equivalence in the contrast thresholds between the laboratory and online repetition conditions, although a substantial amount of data needed to be excluded from the analyses in the latter condition. The contrast threshold was significantly higher in the online non-repetition condition compared with the laboratory and online repetition conditions. To make crowdsourcing more suitable for investigating the contrast threshold, ways to reduce data wastage need to be formulated.

3 **Crowdsourcing visual perception experiments: A case of contrast**

4 **threshold**

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

8 [1] Faculty of Science and Engineering, Waseda University, Tokyo, Japan

9 [2] Faculty of Arts and Science, Kyushu University, Fukuoka, Japan

10 [3]Japan Society for the Promotion of Science, Tokyo, Japan

12 Running head: Crowdsourcing meets Visual Perception

14 *Correspondence:

15 Dr. Yuki Yamada

16 Faculty of Arts and Science, Kyushu University,

17 744 Motooka, Nishi-ku, Fukuoka, 819-0395, Japan.

18 E-mail: yamadayuk@gmail.com

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

23   **Abstract**

24   Crowdsourcing has commonly been used for psychological research but not for studies on sensory

25   perception. A reason is that in online experiments, one cannot ensure that the rigorous settings

26   required for the experimental environment are replicated. The present study examined the

27   suitability of online experiments on basic visual perception, particularly the contrast threshold. We

28   conducted similar visual experiments in the laboratory and online, employing three experimental

29   conditions. The first was a laboratory experiment, where a small sample of participants ($n = 24$;

30   laboratory condition) completed a task with 10 iterations. The other two conditions were online

31   experiments: participants were either presented with a task without repetition of trials ($n = 285$;

32   online non-repetition condition) or one with 10 iterations ($n = 166$; online repetition condition).

33   The results showed significant equivalence in the contrast thresholds between the laboratory and

34   online repetition conditions, although a substantial amount of data needed to be excluded from the

35   analyses in the latter condition. The contrast threshold was significantly higher in the online non-

36   repetition condition compared with the laboratory and online repetition conditions. To make

37   crowdsourcing more suitable for investigating the contrast threshold, ways to reduce data wastage

38   need to be formulated.

40   Keywords: Online experiment, Perception, Vision, Contrast threshold

42   **Introduction**

43   Over the last decade, experiments in psychological research have gone beyond the

44   laboratory. The increasing diversity of research methods and technological advances has increased

45   opportunities for researchers to use resources outside the laboratory. For example, researchers are

46   using outsourcing services to recruit experimental participants and, often, even commissioning

47   research firms to conduct their surveys and experiments. In addition, based on outstanding

48   technological advances in the digital environment and mobile information devices,

49   "crowdsourcing," which is the practice of asking many unspecified people to various kinds of tasks

50   via the internet, has become a powerful tool for psychological research (for a review, see Stewart,

51   Chandler, & Paolacci, 2017).

52   Crowdsourcing can be used for data collection and in asking large numbers of people to

53   participate in surveys or experiments via the internet. Service providers (e.g., Amazon and Yahoo!)

54   manage an experimenter's task and act as a payment agency. The use of crowdsourcing has a

55   number of advantages. The first is its very low cost (e.g., Stewart et al., 2017); for example,

56   participants receive less than 1 USD for responding to a simple questionnaire or engaging in an

57   easy cognitive task. Second, large (more than 1,000 people) and diverse (in age, sex, and culture)

58   samples can easily be employed. The ease in collecting large amounts of diverse data is beneficial

59   not only from the perspective of random sampling but also for planning experiments and

60   estimating the effect size prior to conducting the experiment (Chrabaszcz, Tidwell, & Dougherty,

61   2017). Third, it enables researchers to use their time efficiently. With experiments running all

62   hours of the day and night, data from 1,000 people can be obtained within a day or two, depending

63   on how many active users are registered with the service.

64   Various kinds of online experiments and tasks have been conducted with crowdsourcing.

65   For example, many experimental studies have reported findings based on self-report

66   questionnaires (e.g., Crangle & Kart, 2015; Garcia, Kappas, Küster, & Schweitzer, 2016; Gottlieb

67   & Lombrozo, 2018; Hurling et al., 2017; Sasaki, Ihaya, & Yamada, 2017) and crowdsourced tasks:

68   visual search (de Leeuw & Motz, 2015), reaction time (e.g., Nosek, Banaji, & Greenwald, 2002;

69   Sasaki et al., 2017; Schubert, Murteira, Collins, & Lopes, 2013), keystroke (Pinet et al., 2017),

70   Stroop   (Barnhoorn,   Haasnoot,   Bocanegra,   &   van   Steenbergen,   2014;

71   Crump, McDonnell, & Gureckis, 2013; Majima, 2017), attentional blink (Barnhoorn et al., 2014;

72   Brown et al., 2014), flanker (Simcox & Fiez, 2014; Majima, 2017; Zwaan et al., 2018), Simon

73   (Majima, 2017; Zwaan et al., 2018), lexical decision (Simcox & Fiez, 2014), category learning

74   (Crump et al., 2013), memory (Brown et al., 2014; Zwaan et al., 2018), priming (Zwaan et al.,

75   2018), and decision-making tasks (Berinsky, Huber, & Lenz, 2012; Brown et al., 2014). A

76   previous study using auditory stimuli likewise employed crowdsourcing (Woods, Siegel, Traer, &

77   McDermott, 2017). A recent study recruited infants aged five to eight months via crowdsourcing

78   and measured their looking time with webcams (Tran, Cabral, Patel, & Cusack, 2017). These

79   studies have suggested that the effect size of the performance in such tasks is comparable to that

80   in laboratory experiments; hence, crowdsourcing can be used for diverse online experiments with

81   publishable reliability.

82   However, conventional studies on sensory perception are completed in the laboratory.

83   Moreover, only authors or their laboratory members, who should be well experienced with

84   psychophysical measurements, often participate in experiments on sensory perception. Only a

85   small number of studies have attempted to run sensory perceptual experiments via crowdsourcing.

86   Previous studies have investigated color (Lafer-Sousa, Hermann, & Conway, 2015; Szafir, Stone,

87   & Gleicher, 2014) and randomness (Yamada, 2015) on the web but used one-time color-matching,

88   color word selection, forced choices (same or different), or magnitude estimation tasks. A few

89   studies have measured the point of subjective equality, sensitivity, or thresholds using

90 psychophysical methods in studies on color perception (Ware et al., 2018), volume perception

91 (Pechey et al., 2015), size perception (Brady & Alvarez, 2011) scene perception (Brady, Shafer-

92 Skelton, & Alvarez, 2017), and stimulus visibility (Bang, Shekhar, & Rahnev, 2019). One reason

93 that experiments on sensory perception are rarely conducted online is the necessity for rigorous

94 control over the experimental environment. Online experiments depend significantly on the

95 participant's own computing environment, and experimenters cannot control the display settings,

96 visual distance (or visual field), or lighting conditions. Thus far, online experiments seem

97 unsuitable for experimental studies that focus on the visual functions of spatial and temporal

98 resolutions. For example, in examining the issue of the temporal aspect of stimulus presentation,

99 researchers have found that stimuli are systematically presented for 20 ms longer than the

100 programmed durations (de Leeuw & Motz, 2015; Reimers & Stewart, 2014). However, the above

101 concerns might be negligible, and crowdsourcing is possibly suitable for perception studies. In this

102 case, a large sample could be recruited to bring sufficient statistical power. Further, large and

103 diverse samples are beneficial for the examination of individual differences in perception studies.

104   *Aim of* the present study. This study focused on measuring low-level visual perception via

105 online experiments. We examined the contrast threshold in vision via online crowdsourcing and

106 laboratory experiments. Contrast threshold is a non-temporal visual capacity that is highly

107 susceptible to the influence of the display condition during measurement. Its measurement needs

108 strict linearization of the output of the display with gamma correction; however, most displays of

109 home PCs are not linearized. Moreover, the viewing distance should vary across the participants

110 in the online condition; the spatial frequency depends on this distance. We believed that a

111 comparison between web and lab measurements of visual contrast thresholds would provide

112 tangible evidence of what online experiments can and cannot test regarding non-temporal aspects

113 of stimulus presentation. If the non-linearity of monitor displays, differences in the viewing

114   distance, and other possible factors comprise a negligible random effect, then the contrast threshold

115   online and in the laboratory would be similar. Another important issue is boredom in the

116   participants. In online experiments, boredom in participants substantially decreases data quality

117   (Chandler, Mueller, & Paolacci, 2014); many repetitions are likely to induce boredom. Thus, the

118   present study used two types of iteration for online experiments: the repetition and non-repetition

119   conditions. In the former, participants were presented with each trial 10 times per stimulus

120   condition, whereas in the latter condition, each trial was presented only once. If we could control

121   for measurement errors or individual differences by increasing the sample size, then a single trial

122   for a stimulus condition would suffice to lead to an appropriate conclusion, even in online

123   experiments, without data deterioration. For this reason, the sample size of the participants in the

124   non-repetitive condition was about 10 times that of the repetitive condition.

125

126   **Methods**

127   *Participants*. We used G*Power to determine the sample sizes needed for the repetition condition

128   (α = .05, 1 - β = .80). In the laboratory condition, we used a moderate effect size (f = .25) in the

129   calculation of the required sample size. The required and maximum sample size was 24. In the

130   online repetition condition, we used a small effect size (f = .10) in the calculation of the required

131   sample size, because of the potential for noise in the data from online experiments. The required

132   sample size was revealed to be 138. Considering potential satisficers (Chandler et al., 2014;

133   Oppenheimer, Meyvis, & Davidenko, 2009), who do not devote an appropriate amount of

134   attentional resources to a task and hence cursorily perform it, 200 people was set as the maximum

135   sample size; participants were recruited through a crowdsourcing service (Yahoo! Crowdsourcing:

136   http://crowdsourcing.yahoo.co.jp/). The required sample size in the online non-repetition

137   condition was at least 10 times that in the laboratory condition (240 people) according to the

138   differences in the number of repetitions. Similarly, in the online repetition condition, we recruited

139   300 people as the maximum sample size to account for the potential influence of satisficers. The

140   participants in the laboratory conditions undertook several experiments, including the present

141   experiment, for three hours, and subsequently received 4,000 JPY (the present experiment itself

142   took less than 30 min, although we did not accurately record the duration). The order of these

143   experiments was randomized across the participants. The participants in the online repetition and

144   non-repetition conditions received 50 and 20 T-points (Japanese point service, in which 1 T-point

145   is worth 1 JPY)[1], respectively. The participants were not made aware of the purpose of the study.

146   The experiment was conducted according to the principles laid down in the Helsinki Declaration.

147   The protocol was approved by the ethics committees of Waseda University (approval number:

148   2015-033) and Kyushu University (approval number: 2016-017). We obtained written informed

149   consent from all of the participants in the laboratory condition. Meanwhile, it was difficult to

150   obtain written informed consent in the online conditions. Thus, according to the protocol (approval

151   number: 2016-017), we explained the details of the online experiments in instructions sections,

152   and then asked the participants to take part in the experiments only when they agreed to the

153   instructions. We recruited only PC users to participate in the online experiment.

154   *Apparatus.* In the laboratory condition, stimuli were presented on a 23.5-inch LCD display

155   (FG2421; EIZO, Japan). The resolution of the display was 1920 × 1080 pixels, and the refresh rate

156   was 100 Hz. We performed gamma correction for the luminance emitted from the monitor. The

157   presentation of stimuli and the collection of data were computer-controlled (Mac mini, Apple,

158   USA). We used MATLAB with the Psychtoolbox extension (Brainard, 1997; Pelli, 1997) to

159   generate the stimuli. The observer's visual field was fixed using a chin-and-head rest at a viewing

160   distance of 57 cm from the display. The size information at the visual angle described for the

161   laboratory condition was based on this viewing distance. In the online conditions, the experiment

162   was conducted on a web browser with a JavaScript application (jsPsych; de Leeuw, 2015). jsPsych

163   is a useful toolbox for psychological research, employed in several previous studies (de Leeuw &

164   Motz, 2016; Pinet et al., 2017; Sasaki et al., 2017).

165   *Stimuli and procedure.* Stimuli consisted of a fixation circle (diameter of 0.24 degrees) and Gabor

166   patches, the diameter of which was 42 pixels (2 degrees in the laboratory conditions). The SD of

167   a gaussian function was 6 pixels (0.29°). There were four spatial frequencies of the carrier: 0.02,

168   0.05, 0.09, and 0.38 cycles per pixel (cpp; 0.5, 1, 2, and 8 cycles per degree [cpd] in the laboratory

169   conditions). We set seven contrast levels (the Michelson contrast), varying across the spatial

170   frequencies. The contrasts in the 0.02 cpp (0.5 cpd) trials were 3%, 8%, 13%, 18%, 23%, 28%,

171   and 33%. The contrasts in the 0.05 and 0.09 cpp (1 and 2 cpd) trials were 1%, 6%, 11%, 16%,

172   21%, 26%, and 31%. The contrasts in the 0.38 cpp (8 cpd) trials were 5%, 10%, 15%, 20%, 25%,

173   30%, and 35%. The Gabor patches were tilted 45° clockwise or counterclockwise. We took

174   screenshots of the stimuli on the monitor at the laboratory and then used them for the online

175   conditions.

176   In the laboratory condition, the experiment was conducted in a darkened room. Figure 1

177   shows the timeline of a trial in each of the conditions. The participants initiated each trial by

178   pressing the space key. The fixation circle was presented for 500 ms. After the fixation circle

179   disappeared, the Gabor patch was presented for 50 ms. Then, a blank screen was presented for 300

180   ms, followed by the prompt: "In which direction was the stimulus tilted?" The participants were

181   asked whether the stimulus was tilted clockwise or counterclockwise. They responded without

182   time limits or feedback. Each of the spatial frequency conditions was conducted in a separate

183   session; thus, the experiment consisted of four sessions. The session order was randomized across

184   the participants. In each session, trials were conducted for seven contrasts in two orientations. In

185   the repetition condition, each combination of contrast and orientation was presented 10 times per

186 session. Thus, participants in the repetition condition completed 560 trials, whereas those in the

187 non-repetition condition completed only 56. The order of the trials was also randomized across the

188 participants. Before the first session, we conducted a practice session, in which the participants

189 completed four trials. The spatial frequency of the practice session was identical to that of the first

190 session, and the contrast was 100%. Both of the orientations appeared twice. As in the experiment

191 conditions, the trial order of each session was randomized across the participants.

192 In the online conditions, the participants accessed the page of Crowdsourcing for the link

193 to the web address of the experiment. They navigated to the experiment page via the web address

194 and then input their age and sex. Moreover, after completing the experiment, a four-digit number

195 (8382 and 3599 in the online repetition and non-repetition conditions, respectively) was presented

196 at the final experiment page; the participants typed this number on an empty form on the Yahoo!

197 Crowdsourcing page. The four-digit number was registered for Yahoo! Crowdsourcing in advance.

198 Only when the input and registered numbers corresponded would Yahoo! Crowdsourcing

199 acknowledge that the participants had completed the experiment and give the reward. If the input

200 and registered numbers did not correspond, Yahoo! Crowdsourcing made the participants drop out

201 and did not give the reward. The procedures were identical to that of the laboratory conditions,

202 except for the added insertion of attention check questions (ACQs). This additional step was

203 included because online participants are often distracted (Chandler et al., 2014) or are satisficers

204 (Oppenheimer et al., 2009). ACQs can reduce low-quality responses (Aust, Diedenhofen, Ullrich,

205 & Musch, 2012; Oppenheimer et al., 2009). These tend to be easy calculations based on the four

206 basic arithmetic operations (e.g., 20 + 15 = ?). In the present study, ACQs appeared halfway

207 through the total number of trials in each session and participants selected the correct answer from

208 five options. We conducted the online repetition and online non-repetition conditions from January

209 25 to 28, 2019 and January 29 to February 7, 2019, respectively.

210   *Data analysis.* We excluded participants who gave incorrect answers to one or more of the ACQs.

211   In the laboratory and online repetition conditions, we calculated the contrast threshold of each

212   spatial frequency for each participant, for which the proportion of "correct" responses was 0.82

213   (Cameron, Tai, & Carrasco, 2002; Lee, Baek, Lu, & Mather, 2014), using a probit analysis (i.e.,

214   fitting a cumulative Gaussian function to the proportion of "correct" responses as a function of the

215   contrast level). We used the "glm" function in R (3.4.4). The probit analysis provided the means

216   and standard deviations (SDs) of the distributions. Then, we calculated the contrast thresholds

217   using the means, SDs, and the "qnorm" function in R. We excluded participant data when β

218   calculated by the probit analysis was a negative value. This negative value indicated a reduction

219   in correct responses as the contrast level increased. In such cases, we could not calculate the

220   thresholds. We also excluded the data from participants whose contrast thresholds were less than

221   0 or greater than 100% because the contrast threshold should be within this range. In the online

222   non-repetition condition, we used the pooled data from all the participants and then calculated the

223   contrast threshold for each spatial frequency by the same procedure of the repetition condition.

224   First, to confirm whether the contrast threshold depended on the spatial frequency, we

225   conducted a one-way analysis of variance (ANOVA) on the contrast thresholds, with spatial

226   frequency as a within-participant factor, for the laboratory and online repetition conditions. We set

227   the alpha level at .05 and calculated η_p^2. When the main effects were significant, we conducted

228   multiple comparison tests using Holm's method (Holm, 1979). We conducted the *t*-test six times.

229   Therefore, we increased α from .008 to .05 based on Holm's correction (Holm, 1979).

230   As our purpose was to examine whether the contrast thresholds were different or equivalent

231   between experimental environments in each spatial frequency, we conducted two-tailed Welch's

232   *t*-tests for the contrast thresholds for each spatial frequency. After the *t*-tests, we conducted

233   equivalence tests for the pairs in which the contrast thresholds were not significantly different. For

234   the equivalence tests, we used the TOSTER package in R (Lakens, Scheel, & Isager, 2018) and

235   set Cohen's *d* to 0.5. We compared the contrast threshold of the laboratory condition and the online

236   repetition and non-repetition conditions; thus, we had to conduct *t*-tests and equivalence test three

237   times at most. Therefore, we set α from .017 to .05 based on Holm's correction (Holm, 1979).

238

239   **Results**

240   The results of the proportion of the correct responses and the thresholds in the laboratory and

241   online experiments are shown in Figures 2 and 3, respectively. We collected data from 24 people

242   in the laboratory condition. In the online repetition condition, of the 200 people recruited, only 80

243   participated[1]. As this number did not reach the required sample size, we recruited another 200

244   people and 86 people participated. Hence, we collected data from 166 people in total. For the

245   online non-repetition condition, of the 300 people recruited, only 156 participated. Therefore, we

246   recruited another 250 people and 129 people participated. Hence, we collected data from 285

247   people in total. We excluded the data from 2 (one owing to a negative β and the other, for having

248   a contrast threshold greater than 100%), 84 (53 owing to a negative β; 13, a contrast threshold less

249   than 0; 8, a contrast threshold greater than 100%; and 10, wrong answers to ACQ), and 19 (all

250   owing to wrong answers to ACQ) participants in the laboratory, online repetition, and online non-

251   repetition conditions, respectively, based on the rules detailed in the *Data analysis* section. Thus,

252   we submitted the data from 22 (16 males and 6 females, mean age ± SEM = 21.39 ± 0.39), 82 (54

253   males, 26 females, and 2 non-respondents, mean age ± SEM = 43.56 ± 1.04), and 266 (176 males

254   and 90 females, mean age ± SEM = 42.92 ± 0.61) participants in the laboratory, online repetition,

255   and online non-repetition conditions, respectively, for the statistical analyses.

256   *Effects of spatial frequency within the laboratory and online repetition conditions.* The results of

257   the ANOVA on the contrast thresholds in the laboratory condition revealed that the main effect

258 was significant, $F(3, 63) = 7.63$, $p < .001$, $\eta_p^2 = .27$. The multiple comparison tests showed that the

259 threshold was significantly higher in the 0.5 cpd trials compared with the 1 and 2 cpd trials, $ts(21)$

260 >6.25, $ps < .001$, Cohen's $dzs > 1.33$. Moreover, the threshold was significantly higher in the 4 cpd

261 trials compared with the 2 cpd trials, $t(21) = 2.88$, $p = .009$, Cohen's $dz = 0.61$. The results of the

262 ANOVA on the contrast thresholds in the online repetition condition revealed that the main effect

263 was significant, $F(3, 243) = 26.23$, $p < .001$, $\eta_p^2 = .24$. The multiple comparison tests showed that

264 the threshold was significantly higher in the 4 cpd trials compared with the 1 and 2 cpd trials,

265 $ts(81) > 6.77$, $ps < .001$, Cohen's $dzs > 0.74$. The threshold was also significantly higher in the 0.5

266 cpd trials compared with the 1 and 2 cpd trials, $ts(81) > 4.98$, $ps < .001$, Cohen's $dzs > 0.64$.

267 Moreover, we calculated a McFadden's pseudo R^2 for each of spatial frequency in the laboratory

268 and online repetition conditions and performed the two-way ANOVA on McFadden's pseudo R^2

269 with spatial frequency as a within-participant factor and experimental circumstances as a between-

270 participant factor[2]. As a result, the main effect of spatial frequency was significant ($F(3, 306) =$

271 27.88, $p < .001$, $\eta_p^2 = .21$). Importantly, the main effect of experimental circumstances and

272 interaction were not significant (experimental circumstances: $F(1, 102) = 0.83$, $p = .37$, $\eta_p^2 = .008$;

273 interaction: $F(3, 306) = 0.52$, $p = .67$, $\eta_p^2 = .005$).

274 *Differences and equivalences between laboratory and repeated and non-repeated online*

275 *conditions*. Table 1 shows the summary of the results. For the 0.5 cpd trials, the threshold was

276 significantly higher in the online non-repetition condition compared with the online repetition,

277 $t(332.97) = 6.14$, $p < .001$, Cohen's $d = 0.51$, and laboratory conditions, $t(159.41) = 5.95$, $p < .001$,

278 Cohen's $d = 0.45$. Meanwhile, the online repetition and laboratory conditions showed no

279 significant difference, $t(68.92) = 0.31$, $p = .76$, Cohen's $d = 0.05$. The equivalence test showed

280 significant equivalence between the online repetition and laboratory conditions, $t(68.92) = 2.26$, p

281 $= .013$.

282   For the 1 cpd trials, the threshold was significantly higher in the online non-repetition

283   condition compared with the online repetition, $t(314.58) = 7.54$, $p <.001$, Cohen's $d = 0.55$, and

284   laboratory conditions, $t(285.95) = 7.43$, $p <.001$, Cohen's $d = 0.71$. No significant difference was

285   observed between the online repetition and laboratory conditions, $t(82.43) = 0.56$, $p = .580$,

286   Cohen's $d = 0.09$. The equivalence test showed significant equivalence between the online

287   repetition and laboratory conditions, $t(82.43) = 2.13$, $p = .018$.

288   For the 2 cpd trials, the threshold was significantly higher in the online non-repetition

289   condition than in the online repetition, $t(319.24) = 7.06$, $p <.001$, Cohen's $d = 0.52$, and laboratory

290   conditions, $t(268.92) = 7.11$, $p <.001$, Cohen's $d = 0.72$; no significant difference was found

291   between the online repetition and laboratory conditions, $t(57.31) = 0.33$, $p = .742$, Cohen's $d =$

292   0.06. The equivalence test showed significant equivalence between the online repetition and

293   laboratory conditions, $t(57.31) = 2.12$, $p = .019$.

294   For the 4 cpd trials, the threshold was significantly higher in the online non-repetition

295   condition compared with the online repetition, $t(344.97) = 6.23$, $p <.001$, Cohen's $d = 0.50$, and

296   laboratory conditions, $t(56.41) = 5.06$, $p <.001$, Cohen's $d = 0.51$. However, no significant

297   difference was found between the online repetition and laboratory conditions, $t(31.40) = 0.56$, $p =$

298    $.577$, Cohen's $d = 0.14$. The equivalence test showed that the equivalence between the online

299   repetition and laboratory conditions was marginally significant, $t(31.40) = 1.48$, $p = .075$.

300

301   **Discussion**

302   This study examined whether the contrast threshold was properly measured in an online

303   experiment with two conditions: a condition with repetition of trials and another without repetition.

304   The results showed equivalences in the contrast thresholds of the online repetition and laboratory

305   conditions. The contrast threshold in the online non-repetition condition was higher than that in

306 the online repetition and laboratory conditions. Thus, online experiments seem to be able to

307 measure the contrast threshold as adequately as laboratory experiments, provided enough

308 repetition[3]. Notably, it is difficult to measure contrast thresholds without repetitions. However, as

309 discussed below, there was a high rate of exclusions. In this case, it might be difficult to obtain

310 large and diverse data; thus, one of the advantages of crowdsourcing is possibly lost. Taken

311 together, rash decisions to use crowdsourcing for perception studies is likely to be risky at this

312 time.

313   The present study excluded 51% of the data in the online repetition condition. These

314 exclusions mainly stemmed from the fact that the correct responses decreased as the contrast level

315 increased or the thresholds were under zero. That is, in the online repetition condition, the contrast

316 threshold could be barely calculated precisely. One possibility is that the experimental

317 environment of 49% of the participants in the online repetition condition might be similar to that

318 of the laboratory condition. We were able to calculate the thresholds of these participants and

319 found significant equivalences between the laboratory and online repetition conditions.

320 Meanwhile, the contrast thresholds were much higher in the online non-repetition condition.

321 Although it is difficult to interpret this result, one can argue that the repetitive performance of the

322 experimental task in the online repetition condition caused perceptual learning. It has been well

323 known that contrast discrimination increases with repeated practice or training (e.g., Sowden,

324 Rose, & Davies, 2002; Yu, Klein, & Levi, 2004). However, there are only 10 repetitions for each

325 stimulus in the online repetitive condition, and this little practice does not seem to cause sufficient

326 perceptual learning. Alternatively, the difference in the results with and without repetition may

327 provide clues for problems specific to online experiments. A large amount of the data was excluded

328 in the online repetition condition. Based on this, we can expect the data obtained via online

329 experiments to be noisy. Such noisy data might be included in and mediate the results of the online

330   non-repetition condition. Given the large amount of data exclusion in the online repetition

331   condition and the results of the online non-repetition condition, we could not conclude that online

332   experiments are adequate for measuring the contrast threshold. Indeed, the contrast threshold

333   would be difficult to measure via crowdsourcing unless the lighting conditions of each online

334   participant can be measured and calibrated via camera.

335   There may be solutions for improving the situation of online measurements of the contrast

336   threshold. One would be to control the experimental environment of each participant in the online

337   experiments to match that of a laboratory experiment. A previous study proposed beneficial tips

338   for controlling the size of stimuli, distance from the monitor, sound volume, and brightness

339   (Woods, Velasco, Levitan, Wan, & Spence, 2015). Woods et al. (2015) also provided a possible

340   way to adjust color, which seems to be difficult to control across online participants. They

341   referenced the hints from a psychophysical study (To, Woods, Goldstein, & Peli, 2013) that

342   demonstrated that humans have the ability comparable to a photometer when asked to match two

343   patches in terms of brightness. The potential solution of Woods et al. (2015) was to ask participants

344   to video record their computer screen and a colorful object (reference object) close to the screen

345   using the camera on a mobile device, and then manually calibrate the screen color to the reference

346   object. At this time, these methods require much effort from the participants and experimenters,

347   and prone to technological difficulties; thus, they might not be ultimately effective. The ways to

348   control experimental environments easily should lead to a reduction in low-quality data, and to a

349   decrease in the exclusion of data, while also maintaining the ease of online experiments via

350   crowdsourcing.

351   Another solution is related to participant negligence. In the online experiment, participants

352   might have a difficulty maintaining their motivation while performing tasks; for instance, they

353   may have been unprepared to participate in a psychological experiment and not met the

354   experimenters. Participants with inconsistent motivation often do not devote enough effort to the

355   tasks, and, hence, cursory responses increase (satisficing, e.g., Berinsky, Margolis, & Sances,

356   2016; Maniaci & Rogge, 2014; Miura & Kobayashi, 2016; Oppenheimer et al., 2009). ACQs,

357   which we set during the online condition sessions, are beneficial for protecting the quality of the

358   data from satisficing. It is easy for participants to answer ACQs correctly when they perform the

359   tasks carefully. Generally, it is important to exclude the data from those who wrongly answer

360   ACQs because of inattention and/or cursory responses, to improve the quality of data. However,

361   in the present study, the data exclusion owing to incorrect ACQ responses accounted for 6% of the

362   total data in each of the online conditions. Thus, the ACQ might not have worked as intended in

363   the present study. The type of ACQ was extremely different from that of the main task (i.e., judging

364   the orientation of the Gabor patch). Given this, the ACQ could be improved so that participants

365   are not easily caught out, or another method could be used. An instructional manipulation check

366   (IMC) is also helpful for detecting satisficers (Oppenheimer et al., 2009). An IMC checks whether

367   the participants carefully read the instructions for the tasks. Specifically, they can incorporate the

368   instruction not to answer questions into some methods commonly used in psychological research

369   (e.g., Likert scales); thus, if the participants do not carefully read the instructions, they mistakenly

370   answer the questions. The data from such participants should be excluded because they improperly

371   dealt with the tasks. Additionally, in a recent study, alerting satisficers to their inattentiveness by

372   a repeated IMC was helpful in improving their information processing (Miura & Kobayashi, 2016).

373   In general, ACQs and IMCs are valid tools for the detection and exclusion of data from satisficers.

374   However, it is difficult to prevent satisficers from participating in experiments. To avoid losing

375   data owing to satisficers, blacklisting them might be more effective in the long term.

376   Other ways could be employed to maintain the quality of psychophysical online data. One

377   is developing a platform designed for scientific research. Crowdsourcing services, such as Yahoo!

378 Crowdsourcing and Amazon Mechanical Turk, have some advantages for conducting

379 psychological research. However, they were not developed as research tools and have some

380 inconveniences as well. Recently, a platform for scientific research was designed (TurkPrime,

381 recently rebranded as CloudResearch: Litman, Robinson, & Abberbock, 2017) and integrated with

382 Amazon Mechanical Turk. Prolific is also a remarkable platform for conducting surveys and

383 experiments online (Palan & Schitter, 2018). These helpful systems for improving the quality of

384 online data have also been proposed: Excluding participants based on previous participation,

385 communicating with participants, and monitoring dropout and engagement rates. Elevating these

386 platforms should be helpful for improving the quality of data in online experiments.

387   Contrast sensitivity seemed to be lower in the present study than in the previous ones (e.g.,

388 Cameron et al., 2007; Lee et al., 2014). This discrepancy might be attributed to the intensity level

389 of the stimulus. Several studies have pointed out that the typical hardware used in psychological

390 studies (256 intensity levels, 8 bits) is insufficient for measuring contrast thresholds. One of the

391 solutions is to use a graphics card able to display more than 256 different luminance intensities

392 (e.g., Allard & Faubert, 2008; Lu & Dosher, 2013), but this does not seem to be realistic in online

393 experiments. A previous study proposed the solution of adding visual noise to the stimulus, thereby

394 not requiring special hardware (Allard & Faubert, 2008). This solution might fit the context of

395 online experiments. We aim to address these issues in future studies.

396   Although crowdsourcing does not seem to be suitable for measurements of perception

397 studies at this time, the improvement of environments in online experiments will bring advantages.

398 For example, crowdsourcing enables researchers to obtain large amounts of data from various

399 people, which is advantageous for examining individual differences in perceptual and cognitive

400 processing. In classic laboratory experiments, most participants are university or graduate students,

401 and large amounts of data tend to be difficult to collect. The demographics, personal traits, and

402 cognitive characteristics of the participants do not vary enough to examine the relation between

403 individual differences in perceptual and cognitive processing. Thus, this relation and underlying

404 mechanism have not been understood well, warranting further investigations (Yamada, 2015).

405 Crowdsourcing, however, allows researchers to recruit participants from around the world, and

406 hence, mass data from participants with various personality traits can be collected. Indeed, we and

407 others have already shown the relation between individual differences in personality traits (e.g.,

408 social anxiety, behavioral activation/inhibition systems, and mood) and emotional reactions using

409 crowdsourcing (Chaya et al., 2015; Sasaki et al., 2017). Moreover, we previously conducted a

410 perceptual study indicating the age and sex differences in the perception of pattern randomness

411 (Yamada, 2015). If the environment in online experiments is improved and crowdsourcing

412 becomes suitable for investigating visual perception, then online experiments will be helpful for

413 addressing issues regarding individual differences in visual perception.

414

415 **Conclusions**

416   The present study examined the suitability of online experiments on the contrast threshold.

417 As a result, online experiments seem to be able to measure the contrast threshold as adequately as

418 laboratory experiments, provided enough repetitions. However, there was a high rate of exclusions,

419 which is likely to spoil one of the advantages of crowdsourcing research. Thus, rash decisions to

420 use crowdsourcing for perception studies might be risky at this time. The improvement of

421 technology environments in online experiments via crowdsourcing will bring advantages;

422 individual differences in perceptual processing will be measurable.

424   **Footnotes**

425   [1] Discrepancies between the number of recruitments and that of the actual participants were often

426   found when we used Yahoo! Crowdsourcing. We could not determine the exact reason. One

427   possibility was that the four-digit number was shared online (e.g., SNS) and some crowdworkers

428   may have seen it. In this case, they could be illegally admitted as completing the task by Yahoo!

429   Crowdsourcing even if they did not actually complete the task. Moreover, Yahoo! Crowdsourcing

430   allowed crowdworkers to access the recruiting page only once. Yahoo! Crowdsourcing manages

431   the number of those accessing the recruiting page via Yahoo ID. Crowdworkers, who had multiple

432   Yahoo IDs, could access the recruiting page several times. Therefore, after a participant had

433   completed our experiment, received the four-digit number, and taken the reward, they could access

434   the recruiting page with their other IDs again and input the four-digit number without performing

435   the experiment. These ways of hacking might have caused the discrepancy. Setting and generating

436   unique four-digit number for each participant could prevent this discrepancy; this is impossible at

437   the present system. We plan to discuss means for preventing these issues with Yahoo!

438   Crowdsourcing.

439   [2] We added these post-hoc analyses according to the reviewer's comment.

440   [3] In particular, detection to low contrast stimuli on a non-gamma-corrected monitor are often easier

441   than that on a gamma-corrected monitor: There might be differences in performances for the lowest

442   contrast stimuli between laboratory and online repetition conditions. Thus, according to the

443   reviewer's suggestion, we performed a two-way ANOVA on proportions of the correct responses

444   in the lowest contrast stimuli with spatial frequency (0.5, 1, 2, and 8 cpd) as a within-participant

445   factor and experimental circumstances (laboratory and online repetition) as a between-participant

446   factor. As a result, while the main effect of spatial frequency was significant ($F(3, 306) = 3.34$, p

447   $= .02$, $\eta_p^2 = .03$), the main effect of experimental circumstances and interaction were not significant

448 (experimental circumstances: $F(1, 102) = 0.02$, $p = .89$, $\eta_p^2 < .001$; interaction: $F(3, 306) = 0.27$, p

449 $= .84$, $\eta_p^2 = .003$). Thus, at least, the differences in performances for the lowest contrast stimuli

450 were not found in the present study.

452 **References**

453 Allard, R., & Faubert, J. (2008). The noisy-bit method for digital displays: Converting a 256

454   luminance resolution into a continuous resolution. *Behavior Research Methods*, *40*, 735–

455   743.

456 Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2012). Seriousness checks are useful to

457   improve data validity in online research. *Behavior Research Methods*, *45*(2), 527–535.

458   doi: 10.3758/s13428-012-0265-2

459 Bang, J. W., Shekhar, M., & Rahnev, D. (2019). Sensory noise increases metacognitive efficiency.

460   *Journal of Experimental Psychology: General*, *148*, 437–452. doi: 10.1037/xge0000511

461 Barnhoorn, J. S., Haasnoot, E., Bocanegra, B. R., & van Steenbergen, H. (2014). QRTEngine: An

462   easy solution for running online reaction time experiments using Qualtrics. *Behavior*

463   *Research Methods*, 1–12. doi: 10.3758/s13428-014-0530-7

464 Berinsky, A. J., Huber, G. A., & Lenz, G. S. (2012). Evaluating online labor markets for

465   experimental research: Amazon.com's Mechanical Turk. *Political Analysis*, *20*, 351–368.

466   doi: 10.1093/pan/mpr057

467 Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2016). Can we turn shirkers into workers?

468   *Journal of Experimental Social Psychology*, *66*, 20–28. doi: 10.1016/j.jesp.2015.09.010

469 Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble

470   statistics bias memory for individual items. *Psychological Science*, *22*, 384–392.

471 Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations

472   are critical to rapid scene perception. *Journal of Experimental Psychology: Human*

473   *Perception and Performance*, *43*, 1160–1176.

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

475   10.1163/156856897X00357

476 Brown, H. R., Zeidman, P., Smittenaar, P., Adams, R. A., McNab, F., Rutledge, R. B., & Dolan,

477   R. J. (2014). Crowdsourcing for cognitive science – The utility of smartphones. *PLoS*

478   *ONE, 9* (7), e100662. doi: 10.1371/journal.pone.0100662

479 Cameron, E. L., Tai, J. C., & Carrasco, M. (2002). Covert attention affects the psychometric

480   function of contrast sensitivity. *Vision Research*, *42*, 949–967.

481 Chandler, J., Mueller, P., & Paolacci, G. (2014). Nonnaïveté among Amazon Mechanical Turk

482   workers: consequences and solutions for behavioral researchers. *Behavior Research*

483   *Methods*, *46* (1), 112–130. doi: 10.3758/s13428-013-0365-7

484 Chaya, K., Xue, Y., Uto, Y., Yao, Q., & Yamada, Y. (2016). Fear of eyes: Triadic relation among

485   social anxiety, trypophobia, and discomfort for eye cluster. *PeerJ*, *4*:e1942. doi:

486   10.7717/peerj.1942

487 Chrabaszcz, J. S., Tidwell, J. W., & Dougherty, M. R. (2017). Crowdsourcing prior information

488   to improve study design and data analysis. *PLOS ONE*, *12*: e0188246. doi:

489   10.1371/journal.pone.0188246

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

491   Lawrence Erlbaum.

492 Crangle, C. E., & Kart, J. B. (2015). A questions-based investigation of consumer mental-health

493   information. *PeerJ*, *3*:e867. doi: 10.7717/peerj.867

494 Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical

495   Turk as a tool for experimental behavioral research. *PloS ONE, 8* (3), e57410. doi:

496   10.1371/journal.pone.0057410

497 de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web

498   browser. *Behavior Research Methods*, *47*, 1–12.

499 de Leeuw, J& Motz, B. A. Psychophysics in a Web browser? Comparing response times collected

500   with JavaScript and Psychophysics Toolbox in a visual search task. *Behavior Research*

501   *Methods*, *48*, 1–12.

502   Garcia, D., Kappas, A., Küster, D., & Schweitzer, F. (2016). The dynamics of emotions in online

503   interaction. *Royal Society Open Science*, *3*:8. doi: 10.1098/rsos.160059

504   Gottlieb, S., & Lombrozo, T. (2018). Can Science Explain the Human Mind? Intuitive Judgments

505   About the Limits of Science. *Psychological Science*, *29*, 121–130. doi:

506   10.1177/0956797617722609

507   Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of*

508   *Statistics*, *6*, 65–70. doi: 10.2307/4615733

509   Hurling, R., Murray, P., Tomlin, C., Warner, A., Wilkinson, J., York, G., Linley, P. A., Dovey,

510   H., Hogan, R. A., Maltby, J., & So, T. T. (2017). Short Tips Delivered "in the Moment"

511   Can Boost Positive Emotion. *International Journal of Psychological Studies*, *9*, 88–106.

512   doi: 10.5539/ijps.v9n1p88

513   Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color

514   perception uncovered by "the dress" photograph. *Current Biology*, *25*, R1–R2. doi:

515   10.1016/j.cub.2015.04.053

516   Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for *Psychological*

517   *Research: A Tutorial*. *Advances in Methods and Practices in Psychological Science*, *1*,

518   259–269.

519   Lee, T. H., Baek, J., Lu, Z. L., & Mather, M. (2014). How arousal modulates the visual contrast

520   sensitivity function. *Emotion*, *14*, 978–984.

521   Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing

522   data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*,

523   433–442. doi: 10.3758/s13428-016-0727-z

524   Lu, Z& Dosher, B. (2013). *Visual Psychophysics: From Laboratory to Theory*. MIT Press.

525   Majima, Y. (2017). The feasibility of a Japanese crowdsourcing service for experimental research

526   in psychology. *SAGE Open*, *7*, 1–12.

527   Maniaci, M. R., & Rogge, R. D. (2014). Caring about carelessness: Participant inattention and its

528   effects on research. *Journal of Research in Personality*, *48*, 61–83. doi:

529   10.1016/j.jrp.2013.09.008

530   Miura, A., & Kobayashi, T. (2016). Survey satisficing inflates stereotypical responses in online

531   experiment: The case of immigration study. *Frontiers in Psychology*, *7*:1563. doi:

532   10.3389/fpsyg.2016.01563

533   Nosek, B. A., Banaji, M., & Greenwald, A. G. (2002). Harvesting implicit group attitudes and

534   beliefs from a demonstration web site. *Group Dynamics: Theory*, *6* (1), 101–115. doi:

535   10.1037/1089-2699.6.1.101

536   Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks:

537   Detecting satisficing to increase statistical power. *Journal of Experimental Social*

538   *Psychology*, *45* (4), 867–872. doi: 10.1016/j.jesp.2009.03.009

539   Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of*

540   *Behavioral and Experimental Finance*, *17*, 22–27.

541   Pechey, R., Attwood, A. S., Couturier, D. L., Munafò, M. R., Scott-Samuel, N. E., Woods, A., &

542   Marteau, T. M. (2015). Does glass size and shape influence judgements of the volume of

543   wine? *PLOS ONE*, *10*: e0144536.

544   Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers

545   into movies. *Spatial Vision*, *10*, 437-442. doi: 10.1163/156856897X00366

546   Pinet, S., Zielinski, C., Mathot, S., Dufau, S., Alario, F. X., & Longcamp, M. (2017). Measuring

547   sequences of keystrokes with jsPsych: Reliability of response times and inter-keystroke

548   intervals. *Behavior Research Methods*, *49*, 1163–1176.

549   Plant, R. R., & Turner, G. (2009). Millisecond precision psychological research in a world of

550   commodity computers: new hardware, new problems? *Behavior Research Methods*, *41*

551   (3), 598–614. doi: 10.3758/s13428-016-0776-3

552   Reimers, S., & Stewart, N. (2015). Presentation and response timing accuracy in Adobe Flash and

553   HTML5/JavaScript Web experiments. *Behavior Research Methods*, *47*, 309–327. doi:

554   10.1016/j.chb.2015.12.049

555   Sasaki, K., Ihaya, K., & Yamada, Y. (2017). Avoidance of novelty contributes to the uncanny

556   valley. *Frontiers in Psychology*, *8:* 1792. doi: 10.3389/fpsyg.2017.01792

557   Schubert, T. W., Murteira, C., Collins, E. C., & Lopes, D. (2013). ScriptingRT: A Software Library

558   for Collecting Response Latencies in Online Studies of Cognition. *PLOS ONE*, *8* (6),

559   e67769. doi: 10.1371/journal.pone.0067769

560   Simcox, T., & Fiez, J. A. (2014). Collecting response times using Amazon Mechanical Turk and

561   Adobe Flash. *Behavior Research Methods*, *46*, 95–111. doi: 10.3758/s13428-013-0345-

562   y

563   Sowden, P. T., Rose, D., & Davies, I. R. L. (2002). Perceptual learning of luminance contrast

564   detection: Specific for spatial frequency and retinal location but not orientation. *Vision*

565   *Research*, *42*, 1249–1258. http://doi.org/10.1016/S0042-6989(02)00019-6

566   Stewart, N., Chandler, J., & Paolacci, G. (2017). Crowdsourcing Samples in Cognitive Science.

567   *Trends in Cognitive Sciences*, *21*, 736–748. doi: 10.1016/j.tics.2017.06.007

568   Szafir, D. A., Stone, M., & Gleicher, M. (2014). Adapting color difference for design. *Proceedings*

569   *of Color and Imaging Conference*, *2014*, 228–233.

570   To, L., Woods, R. L., Goldstein, R. B., & Peli, E. (2013). Psychophysical contrast calibration.

571   *Vision Research*, *90*, 15–24. doi: 10.1016/j.visres.2013.04.011

572   Tran, M., Cabral, L., Patel, R., & Cusack, R. (2017). Online recruitment and testing of infants with

573   Mechanical Turk. *Journal of Experimental Child Psychology*, *156*, 168–178. doi:

574   10.1016/j.jecp.2016.12.003

575   Turpin, A., Lawson, D. J., & McKendrick, A. M. (2014). PsyPad: a platform for visual

576   psychophysics on the iPad. *Journal of Vision*, *14* (3):16, 1–7. doi: 10.1167/14.3.16

577   Ware, C., Turton, T. L., Bujack, R., Samsel, F., Shrivastava, P., & Rogers, D. H. (2018). Measuring

578   and Modeling the Feature Detection Threshold Functions of Colormaps. *IEEE*

579   *transactions on visualization and computer graphics*. doi: 10.1109/TVCG.2018.2855742

580   Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to

581   facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*,

582   2064–2072.

583   Woods, A. T., Velasco, C., Levitan, C. A., Wan, X., & Spence, C. (2015). Conducting perception

584   research over the internet: A tutorial review. *PeerJ*, *3*:e1058. doi: 10.7717/peerj.1058

585   Yamada, Y. (2015). Gender and age differences in visual perception of pattern randomness.

586   *Science Postprint*, *1* (2): e00041. doi: 10.14340/spp.2015.01A0002

587   Yu, C., Klein, S. A., & Levi, D. M. (2004). Perceptual learning in contrast discrimination and the

588   (minimal) role of context. *Journal of Vision*, *4* (3), 4–14. http://doi.org/10.1167/4.3.4

589   Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., &

590   Zeelenberg, R. (2018). Participant nonnaiveté and the reproducibility of cognitive

591   psychology. *Psychonomic Bulletin & Review*, 25, 1968–1972. doi: 10.3758/s13423-017-

592   1348-y

593

595 **Acknowledgments**

596 We would like to thank Dr. Daiichiro Kuroki for developing the program of the online experiment.

597 This research was supported by JSPS KAKENHI #17J05236 to K.S. and #15H05709, #16H03079,

598 #16H01866, #17H00875, #18H04199, and #18K12015 to Y.Y.

599

600   **Competing interests**

601   The authors declare no competing interests.

602

603 **Author contributions**

604 Contributed to conception and design: KS, YY

605 Contributed to acquisition of data: KS

606 Contributed to analysis and interpretation of data: KS

607 Drafted and/or revised the article: KS, YY

608 Approved the submitted version for publication: KS, YY

609

610   **Data accessibility statement**

611   The   dataset   and   the   R   codes   for   the   analyses   are   shown   in

612   https://figshare.com/s/c067967c1cfd3238244b

613

614

615

616   *Figure legends*

617   Figure 1. Timeline of a trial in all the conditions. For enhanced visibility, we presented the stimulus

618   in 100% contrast level in this figure.

619   Figure 2. Results of the correct responses in the laboratory and online experiments.

620   Figure 3. Results of the thresholds in the laboratory and online experiments. Error bars denote

621   standard deviations.

622

623   *Table legends*

624   Table 1. Summary of the results in differences and equivalences between laboratory, repeated, and

625   non-repeated online conditions. Note: NHST 95% CI = Null Hypothesis Significant Test 95%

626   confidence interval, for cases of a significant difference between pairs; TOST 90 % CI = Two

627   One-Sided Test 90% confidence interval, for cases of a (marginally) significant equivalence

628   between pairs.

629

Figure 1

Timeline of a trial in all the conditions

For enhanced visibility, we presented the stimulus in 100% contrast level in this figure.

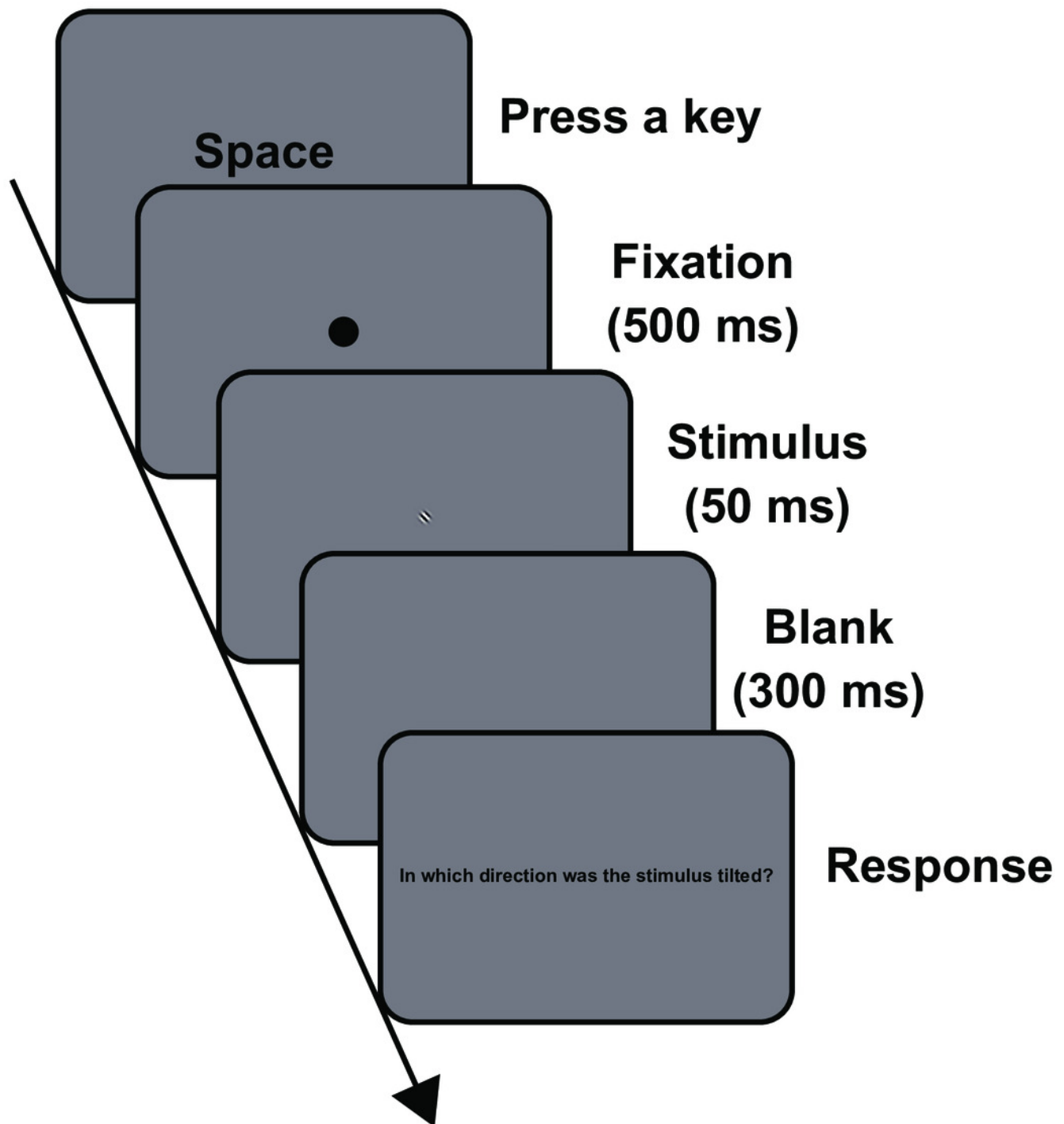



Figure 2

Microsoft Word - RevMS3rd_CrowdContrast_20191121_notracked.docx Results of the correct responses in the laboratory and online experiments.

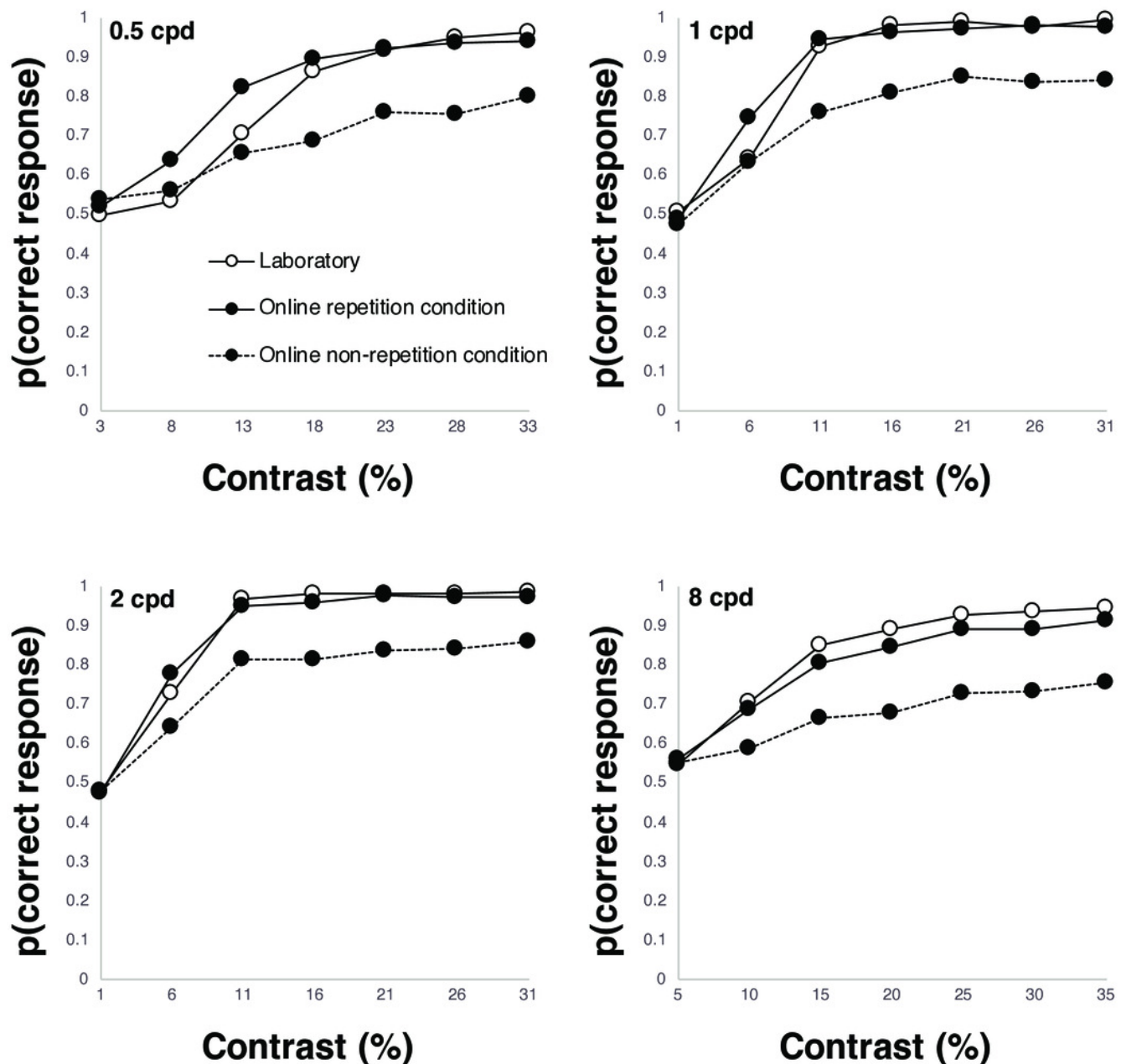
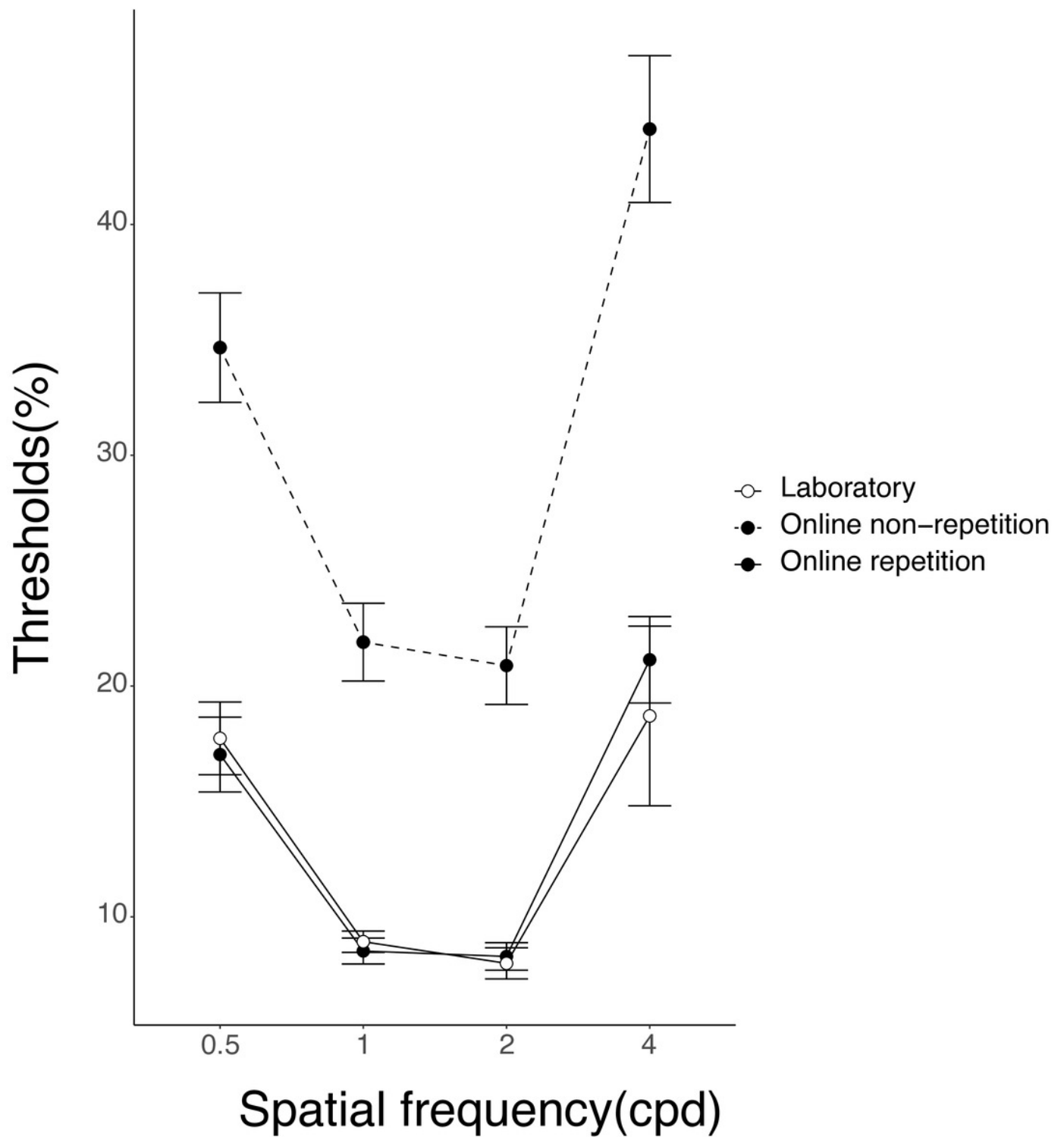



Figure 3

Results of the thresholds in the laboratory and online experiments.

Error bars denote standard deviations.

Summary of the results in differences and equivalences between laboratory, repeated, and non-repeated online conditions.

NHST 95% CI = Null Hypothesis Significant Test 95% confidence interval, for cases of a significant difference between pairs; TOST 90 % CI = Two One-Sided Test 90% confidence interval, for cases of a (marginally) significant equivalence between pairs.

cpd		Laboratory	Online repetition	Online non-repetition
0.5	Laboratory			
	Online repetition	**Sig. Eq.** TOST 90% CI: -3.1 − 4.5		
	Online non-repetition	**Sig. Dif.** NHST 95% CI: -22.6 − -11.3	**Sig. Dif.** NHST 95% CI: -23.3 − -12.0	
1	Laboratory			
	Online repetition	**Sig. Eq.** TOST 90% CI: -0.8 − 1.6		
	Online non-repetition	**Sig. Dif.** NHST 95% CI: -16.4 − -9.5	**Sig. Dif.** NHST 95% CI: -16.9 − -9.9	
2	Laboratory			
	Online repetition	**Sig. Eq.** TOST 90% CI: -1.8 − 1.2		
	Online non-repetition	**Sig. Dif.** NHST 95% CI: -16.5 − -9.3	**Sig. Dif.** NHST 95% CI: -16.1 − -9.1	
4	Laboratory			
	Online repetition	**Marg. Sig. Eq.** TOST 90% CI: -9.8 − 4.9		
	Online non-repetition	**Sig. Dif.** NHST 95% CI: -35.5 − -15.4	**Sig. Dif.** NHST 95% CI: -30.3 − -15.7	