Summary of the paper and generel comments

This manuscript reported the empirical study that investigated the viability of online experiments with crowdsourcing population in the domain of human sensory perception. Many empirical studies in psychology and other social sciences have recently used the crowdsourcing service as a measure for data collection, however, it remains unclear whether the online experiment is also a powerful tool for studies in relatively low-level sensory perception, since these experiments require more precise control of physical environment (e.g., lighiting, apparatus, …) compared to studies of higher cognition, such as thinking. The authors conducted an experiment that compared visual contrast threshold between participants from conventional student sample and online crowdsourcing sample. Results indicated that data from the conventional laboratory experiment and the online experiment with the same number of iterations of trial as laboratory are equivalent, however data from online without repetition indicated significantly greater contrast threshold than other conditions. Results also exhibited that the data from online experiment may suffer from a high level of data exclusion partly due to lack of experimental control in the presentation of visual stimulus. From these results, the authors (seemed to) conclude that the online crowdsourcing experiment is useful tool for collecting data even in the visual perceptual study, and they also pointed out practical tips on conducting crowdsourced studies in this field.

The empirical study reported here is well-designed and results from the experiment seems to be beneficial for researchers studying human perceptual process. Unfortunately however, the manuscript as it is, failed to make an important contribution in the literature for two reasons. First of all, it seems to me that the central claim of the manuscript failed to show its distinctness. Second, procedures of the experiment (mainly about sampling and data exclusion) and displays of results cannot be understood easily, therefore should be revised accordingly. In the following, I will describe problems of the manuscript in detail.

Major problem

The central claim of the manuscript

In the present manuscript, the authors' central claim was not explained clearly. I understood that the primary aims of present study is to exhibit the usefulness of crowdsourced online experiments for studies of low-level sensory perception. However, throughout the manuscript, the authors discussed about many cons of crowdsourcing (abbreviated to CS, hereafter) usage, such as the lack of control in environment, considerable amount of missing data. On the other hand, they did not discuss much about the pros of CS. In fact, the results of the present experiment may suggest that the CS experiments on the low-level visual perception cannot be recommended due to its high exclusion rate. However, in the discussion section, the authors seemed to conclude that CS is a profitable tool for perception studies

even if researchers should take into consideration of its high rate of exclusion. If the authors intended to argue benefits of CS, they might want to describe pros of CS more in detail throughout the entire manuscript. On the other hand, if they intended to give a warning to rash decisions to use CS in low-level perceptual studies, the discussion section should be rewritten accordingly.

It is also unclear the necessity of online CS studies in the domain of low-level perceptual process, since experiments in this domain often require relatively small sample size compared to those investigating higher level of cognition. Of course, there may be several advantages in the online perceptual studies. However, it seems to me that high demands for (and costs of) precise control of experimental environment in perceptual research cancel out the benefits of CS. If the authors, nevertheless, want to claim the CS as an useful tool for perceptual study, they ought to discuss throughly about advantages over conventional methods.

In the followig, I will also describe the poiny-by-point comments related to this issue.

- p.4, l. 18,
  I would recommend that authors describe the importance (or the pros) of conducting online experiments investigating low-level visual perception. It seems to me, in the previous paragraph, that your arguments are mostly the cons of online experiments of sensory perception, therefore, readers might misunderstand the purpose of the present study.
- p.5, l. 2, non-linearity of monitor display
  If you thought that the non-linearity of display is the main issue, you probably don't need to compare highly controlled laboratory environment and uncontrolled (and noisy) CS environment directly, rather you can say something if you compare the data from linearlized display (located in a darkroom and gamma-corrected) and those from non-linearized display (located in a regular laboratory room without any correction). It seems that the latter comparison might be suitable in order to address the non-linearity issue. Regarding this, it seems unclear the reason why you used different software in between the laboratory condition (Matlab+Psychtoolbox) and the online condition (jsPsych+Web browser). Is there any possibility that the 'less controlled' laboratory setting (e.g., the monitor is located in a room with natural light, using jsPsych and Web browser) results in any difference compared to controlled laboratory setting? It seems to me that the authors might need to compare the data from the above less-controlled laboratory participants with those from controlled participants. I believe that the authors ought to provide evidence, at least, showing that the difference in the software doesn't affect results when the experiments has been conducted in the same environment; or to discuss this issue as an unresolved question in the general discussion.
- p.5, l.4, the present study used two types of iteration.
  When I firstly came to this sentence, I could not understand why you manipulate iterations in this experiment.

- p.13, l. 12-14.
  Increasing a control in experimental environment in CS may help to acheive high-quality data (and reduce data exclusion as a result), however, at the same time, it will also hamper the important aspect of online studies, namely effortlessness in conducting empirical studies.
- p. 15, l.8, l. 11 personal (personality) traits.
  It seems that the authors did not discuss about any specific personality traits here. It will advisable if you specify what kind of personality traits are you discussing here, and why introducing CS in perceptual study may boost a diversity in the collected sample. In addition, the authors might want to discuss about the diversity not only in terms of personality traits, but also other aspects of individual differences (e.g., demographic, cognitive characteristics).

Sampling issue

I would suggest the authors describe the criteria of data exclusion and the number of participants (or trials) excluded due to these criteria in detail.

At first, the authors described the required sample size of three conditions in the 'Participants' subsection, however they also said they recruited additional participants for two online conditions due to data exclusion in the beginning of the Results section. Readers cannot understand how many participants (or trials) remained in the final sample when they looked at the Methods section. In addition, the description about data exclusion in the Results is quite confusing. For example, the authors said 'of the 200 people recruited, only 80 participate (p. 10, l.1)', but I cannot understand why the other 120 recruited people did not 'participate' in the experiment (the same comment is also applied to the non-repetition condition). The authors should describe the criterion of this exclusion. The authors also said that they recruited additional 200 participants and 'collected data from 166 participants (p.10, l.3)'. This sentence is quite confusing because it is unclear that these 166 participants were drawn from either the additional 200 people or the total 400 people.

Furthermore, the authors said they excluded another 105 participants from analyses based on the rules described in the Methods. How the criteria for exclusion here differs from the previous? It would be also advisable if the authors provide detailed information about exclusion (i.e. the number of excluded participants for each rule) since they adopted multiple rules for exclusion (negative beta and out-of-range threshold).

Displaying the results

The results of the present experiments seems straightforward, however I believe there's room for improvement for better understanding. I would suggest that…

- p.9, l. 24, Figure 1
  Error bars are not easily distinguishable.

- p.10, l. 9, the main effect was significant
  It would be advisable if you could mention the IV here.
- p.11, l.2-
  It would be advisable if the authors summarize the results of difference/equivalence tests in the table with useful statistics, such as equivalance interval (test boundaries), mean difference (and its confidence interval).

In addition to the major issues described here, I would also point out several minor issues followed by grammatical mistakes.

Minor problem

Definition of crowdsourcing

It seems to me that the usage of the term 'crowdsourcing' in the present manuscript is slightly different from its definition, in other words, the term is inaccurately defined.

- p.2, l. 10, Crowdsourcing is …
  By original definition, CS is not the method of recruiting research participants, although it "can be used as a measure of data collection".
- p.2 l. 21, In brief, …
  If you intended to summerize the content of this paragraph in this sentence, you should also mention the advantages of CS other than its efficacy. If you intended to summerize the previous sentence, you have already mentioned the efficacy of data collection with CS, therefore this sentence seems redundant.
- p.2 l. 23, crowdsourcing has been used for various kinds of experiments and tasks.
  I thought that CS itself is a tool for participants recruitment but is not for the tool for conducting (controlling, more precisely) online experiments.

The title of subsection

In the Results section, the title of two subsections failed to express its contents adequately. I would use 'Effects of spatial frequency within each experimental condition' as the former and 'Differences and equivalences between laboratory, repeated and non-repeated online conditions' as the latter.

other issues

- p.3, l. 17, The high reliability of some experimental data was also confirmed.
  You might want to describe what was confirmed more in detail. In the end of previous sentence, you argued that data from CS were not reliable in some tasks (e.g. subliminal priming). But, in the

last sentence of this paragraph, you argued that CS has successfully shown its reliability. It would strengthen the advantage of CS, if you describe what kind of phenomena have been shown (or replicated) with high reliability with CS.

- p.3, l. 22, it is not unusual that…
  It seems unclear what 'unusual' means here. Is it because the experiment 'usually' recruited only quite a few participants who have plenty of experience in psychophysical experiments?
- p.4, l. 14, functions of spatial and temporal resolution such as contrast threshold
  Is the contrast threshold temporal visual function? In the next parapragh, you argued that 'the contrast threshold is a non-temporal visual capacity'.
- p.6, l. 7, It would be advisable if you could explain what 'T-points' mean more in detail.
- p.7, Stimuli and Procedure section
  It would advisable if you provide the figure of sample stimulus.
- p.12, l.2-3, On the other hand, at high spatial frequencies, …
  It seems that the contrast threshold in non-repetition condition is always (I mean, not specifict to high spatial frequencies) higher than the other two conditions.
- p.12, l.9-,
  Is there any possibility that differences in demographics (such as age) might affect the present results? You might want to discuss about the possibility. In addition, demographic characteristics of online participants should be described in the Methods section.
- p.13, l. 7-8,
  You cited Woods et al (2015)'s study to discuss about possible improvement of the online perceptual study but did not explain the tips in detail. It would be advisable if you could describe specific tips that are important to visual perception study.
- p.14, l. 13, TurkPrime
  I kind of remember that TurkPrime has been rebranded as CloudResearch. Therefore, it will be advisable if you mention this information. In addition, Prolific (https://www.prolific.co/) is also a good alternative to conduct surveys and experiments in online. FYI, please refer, Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. Journal of Behavioral and Experimental Finance, 17, 22-27. doi:10.1016/j.jbef.2017.12.004

Grammatical issues

Abstract

- p.1, l.13, although a high level of data exclusion .. repetition condition
  I would say 'although a substantial amount of data needed to be excluded from anlaysis in the online repetition condition' here.

Introduction

- p. 4, l. 13, online experiments seem inadequate
  It seems to me that ‘unsuitable’ might fit better in this sentence.

Methods

- p. 8, l. 24, precisely
  It seems that ‘correctly’ might be more natural here. Or, the last phrase of this sentence (, and in this case, … the task precisely) may not be needed.
- p. 9, l. 1, the data of participants => ‘the data from participants’ would be better here.
- p. 9. l. 2, under 0 or over 100% => ‘less than 0 or greater than 100%’

Discussion

- p. 12, l. 19, repetition condition may inform our understanding.
  I cannot understand what you are trying to say here.
- p. 12, l. 21, noisy data should be included => ‘might be included’ seems to be better here.
- p. 12, l. 23, Considering the exclusion of data
  I would say ‘Considering a large amount of data exclusion’ here.
- p. 13, l. 17, did not face experimenters
  ‘did not meet experimenters in person’ seems to be suitable here.

I would also suggest that the manuscript is proofed by native English speaker since many expressions seems to be unnatural and hinder smooth understanding of the argument.