**Revision**

We thank the editor, as well as the reviewer for the time taken to review our work, and for the encouraging comments and insightful remarks. We have answered, and addressed, all points below. All corresponding corrections are bolden in the main manuscript document. Major restructuring of text and changes are highlighted in yellow. The additions/modifications (underlined) in the tables attached in the manuscript are also reflected in the individual editable tables uploaded.

**REVIEWER #1**

1. Literature.

The introduction mentions the evidence against a general factor. This part is critical for the aim of the paper, and needs to be expanded. The evidence should be described, as then the reader can compare to the new evidence. Differences in methodology are important. A few more recent papers are missing:

Cretenoud, et al. (2021) Individual differences in the perception of visual illusions are stable across eyes, time, and measurement methods. Journal of Vision 21(5):26

Cretenoud, et al. (2021). How do visual skills relate to action video game performance?. Journal of Vision, 21(7), 10.

Cretenoud, et al. (2020). Individual Differences in the Muller-Lyer and Ponzo Illusions Are Stable Across Different Contexts. Journal of Vision, 20(6):4, 1-14.

We agree that the literature review in our introduction could be expanded upon and thank the reviewers for the suggested papers. We have since incorporated the relevant evidence from these papers accordingly in lines (XXX):

*Moreover, conventional paradigms often focus on the participants' subjective experience, by asking them the extent to which they perceive two identical targets as different26, having them estimate the targets' physical properties27, or most typically, through a method of adjustment, having them adjust the targets to perceptually match a reference stimulus16,28,29,30.*

*While some recent studies have made efforts to implement paradigms that are more empirically rigorous31, varying the physical dimensions of the elements that make up the visual illusions, most of the manipulations made only alter the illusion magnitude without modulating the actual strength of the illusory effect.*

2. Stimuli and procedure.

The description of the stimuli is not sufficient, there is too much reliance on linking to external sources. In the paper the illusions should be listed and described carefully one by one. From the list it seems that there are at least two types, some are old optical-geometrical illusions (Delboeuf, Ebbinghaus, Vertical-Horizontal, Zöllner, Müller-Lyer, Ponzo, Poggendorff) while at least two have to do with brightness (White, Contrast). I assume contrast refers to simultaneous brightness contrast.

TODO. Anshu

Brightness illusions are highly depended on the display used, and not ideal for online studies. It may be worth analysing these two sets separately.

We thank the reviewer for this suggestion. It is true that contrast-based illusions are somewhat different from the others, likely relying on different perceptual processes. To avoid any assumptions about potential groupings of illusions, we have indeed started by analyzing them separately (i.e., the sensitivity scores were extracted for each illusion type separately). We have clarified that in the data analysis section (lines XX):

*The first part of the analysis focused on modelling the effect of illusion strength and task difficulty on errors and response time (RT) separately for each illusion.*

The combination of the scores from different illusions only takes place at the factor structure exploration phase. We inspected a bit further these results (available at <https://realitybending.github.io/IllusionGameValidation/study2/study2.html>) to make sure we did not miss anything related to that. First of all, the correlation between the scores from the two contrast-based illusions does not seem abnormally higher as compared to the rest of the values (the effect of illusion strength for the Contrast illusion does correlate the strongest with the strength effect for White, r = .290, but the magnitude does not stand out in the general pattern).

Chart

Description automatically generated

Second, we did indeed a Structural Equation Model (m2) that grouped the two contrast-illusions separetely:

Lines =~ MullerLyer\_Strength + MullerLyer\_Interaction + Ponzo\_Strength + Ponzo\_Interaction + VerticalHorizontal\_Strength + VerticalHorizontal\_Interaction

Circles =~ Ebbinghaus\_Strength + Ebbinghaus\_Interaction + Delboeuf\_Strength + Delboeuf\_Interaction

Contrast =~ Contrast\_Strength + Contrast\_Interaction + White\_Strength + White\_Interaction

Angle =~ Zollner\_Strength + Zollner\_Interaction + Poggendorff\_Strength + Poggendorff\_Interaction + RodFrame\_Strength + RodFrame\_Interaction

i =~ Lines + Circles + Contrast + Angle

But this model did not outperform more parsimonious and straightforward models.

anova(m0, m1, m1b, m2, m4)

*## Chi-Squared Difference Test*

*##*

*## Df AIC BIC Chisq Chisq diff Df diff Pr(>Chisq)*

*## m1 160 11299 11475 399*

*## m1b 160 11299 11475 399 0 0*

*## m0 163 11483 11648 589 190 3 <2e-16 \*\*\**

***## m2 166 12479 12633 1590 1002 3 <2e-16 \*\*\****

*## m4 170 13378 13519 2498 907 4 <2e-16 \*\*\**

It seems like even though contrast-based illusions might share different perceptual process at a low perceptual level, the inter-individual variability in their sensitivity is partly related to that of other illusions.

3. Terminology

Terminology is important, and especially in the case of illusions. I will take a specific sentence to discuss a problem:

"illusions are conceptualized as ambiguous percepts (noisy sensory evidence) giving ample weight to prior knowledge to minimize prediction error and provide a coherent perceptual experience"

In most illusions the percept is not ambiguous at all, as the percept refers to the experience of the observer. What can be described as ambiguous is the stimulus.

TODO. Anshu?

4. Results

The presentation of the results also needs a complete restructuring. Most of the results are summarised in Figure 3. This Figure is very busy, confusing, and hard to understand. There are too many panels, and they are not labelled. Some are summary of raw data, some are results from model fitting. Many axes are not labelled.

There is enough material here for at least three separate figures. This way basic summary statistics (including scatter graphs) can be presented first, and the model introduced only later.

TODO. Dom. Perhaps split figure. Drop demographic data.

5. Large number of trials

Each participant makes 1340 responses (trials). I am worried that over such a long experiment (online) some different strategies may be adopted. Some people will remain motivated while others will "cheat". The problem is that these different strategies are hard to control online and they are likely to correlate with personality traits. The authors should at the very least consider this possibility.

We agree with the reviewers that problematic trials are particularly likely given the nature of the experiment (many trials + online study). First of all, the crowdsourcing platform used, **prolific**, is considered as the one providing the most reliable data (see Peer et al., (2022). Data quality of platforms and panels for online behavioral research. *Behavior Research Methods, 54*(4), 1643-1662.). As it is possible to flag “bad” participants by the experimenters (a score that is tracked by the platform which, I believe, would filter out consistently problematic participants), they are motivated to do the tasks thoroughly (also avoiding this way possible rejections).

That being said, we have further adopted several steps to ensure data quality, detailed below:

Regarding motivation given the number of trials:

* **Management of expectations**: The total length was … and participants were (hopefully) ready and prepared to spend that time in front of a computer.
* **Gamification**: we tried to “gamify” our experiment, keeping it dynamic and user-friendly (for instance, a completion progress bar was shown in between blocks, giving participants a sense of their progression). Though this a purely subjective feeling (and naturally biased), it seems to us that the nature of the stimuli makes for a somewhat fairly engaging experience (maybe not as far as “fun”, but one goes from one trial to the next without dread). The experiment can be tested at https://realitybending.github.io/IllusionGameValidation/study2/index.html.
* **Total length**: The relatively short ISI allowed for a fast trial succession, which helped make the experiment more dynamic (“lively”).
* **Multiple phases**: Finally, the trials were divided into 2 sets of 10 illusion blocks, each defined by a specific illusion type. Again, this increased the impression of progression and allowed the participants to take short mental breaks when they desired so. Moreover, the 2 sets were further separated by 2 short personality questionnaires (Mini IPIP-6 and PID-5), attenuating any accumulated effects of fatigue from completing the illusion blocks.

Regarding outlier trials:

* **Block rejection**: we carefully inspected the reaction time distribution for each participant separately, for each block. See figure <https://realitybending.github.io/IllusionGameValidation/study2/index.html> (note that the figure is quite big and takes a few second to load). As the RT distribution typically follows a well-defined log-normal distribution, blocks or participants that had an abnormal distribution were quite salient (see distributions colored in yellow or red). One of the common scenario (often in yellow) was participants which first block is okay, but which second block is collapsed to the left (i.e., very short RTs), as if they completed the first half of the experiment well, but when instructed to go back doing the task after doing the personality questionnaires, they just entered their answer as fast as possible.
* **Participant rejection**: we also computed the average RT and error rate per participant (see descriptive table at <https://realitybending.github.io/IllusionGameValidation/study2/study2.html#exclusions>) to help us identify outliers.
* **Trial-wise outliers**: Finally, within the remaining data, we removed trials with a trailing RTs relative to each participant’s mean RT (see top figure in <https://github.com/RealityBending/IllusionGameValidation/blob/main/study2/figures/outliers_RTtrials-1.png>).

**REVIEWER #2**

This paper examines whether illusion susceptibility can be parametrically varied in a range of illusions and in a large sample. Furthermore it examines if generalized illusion susceptibility is influenced by demographics and personality traits. This paper has several strengths, including preregistration, a number of illusions, a large sample, open materials and very nice data visualization. The main area for improvement I believe are in the analysis section, where the use of such a large number of tests could be more clearly justified and explained in a way that makes the analyses easier to navigate given the hypotheses and aims.

Signed: Rebecca Hirst

We sincerely thank Dr Hirst for her thorough review, and are truly pleased that she appreciated (at least some parts of) our study.

There are a lot of different statistical models (GAMS, Bayesian Logistic models, General linear mixed models, EFA, SEM) and it could be more clearly justified why each is needed

TODO.

and what set of transformations were used and why (e.g. line 196 – 203 indicate that different scores were transformed in different ways, log sqrt, cbrt

TODO.

please can the authors outline the approach taken and models that were compared).

TODO.

The pre-registration only mentions Bayesian models and, whilst straying from pre-registration is fine, a clear justification in the text would be useful and help the reader to navigate through all of these tests and what each is doing.

TODO.

There also needs to be clearer explanation of the EFA an SEM models to allow replication. Indeed, is the EFA needed since the final SEM selected holds all illusions as independent, loading onto a single factor?

TODO. Dom.

Line 196: what is “log(diff)”? Please define “diff” here.

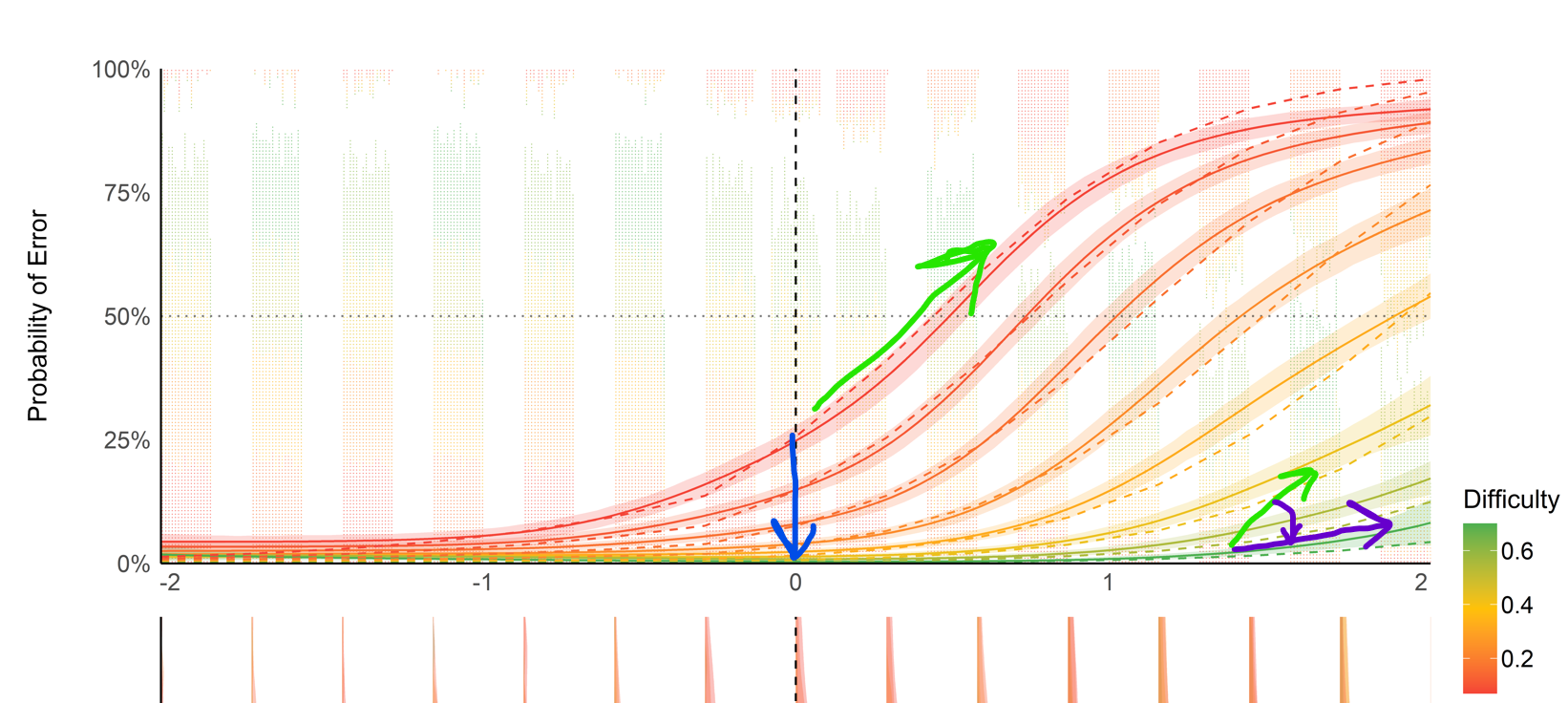
TODO. Anshu.

Figure 2 – it looks as though the interaction with task difficulty loads positively for some illusions and negatively for others – might the authors comment on this difference between tasks?

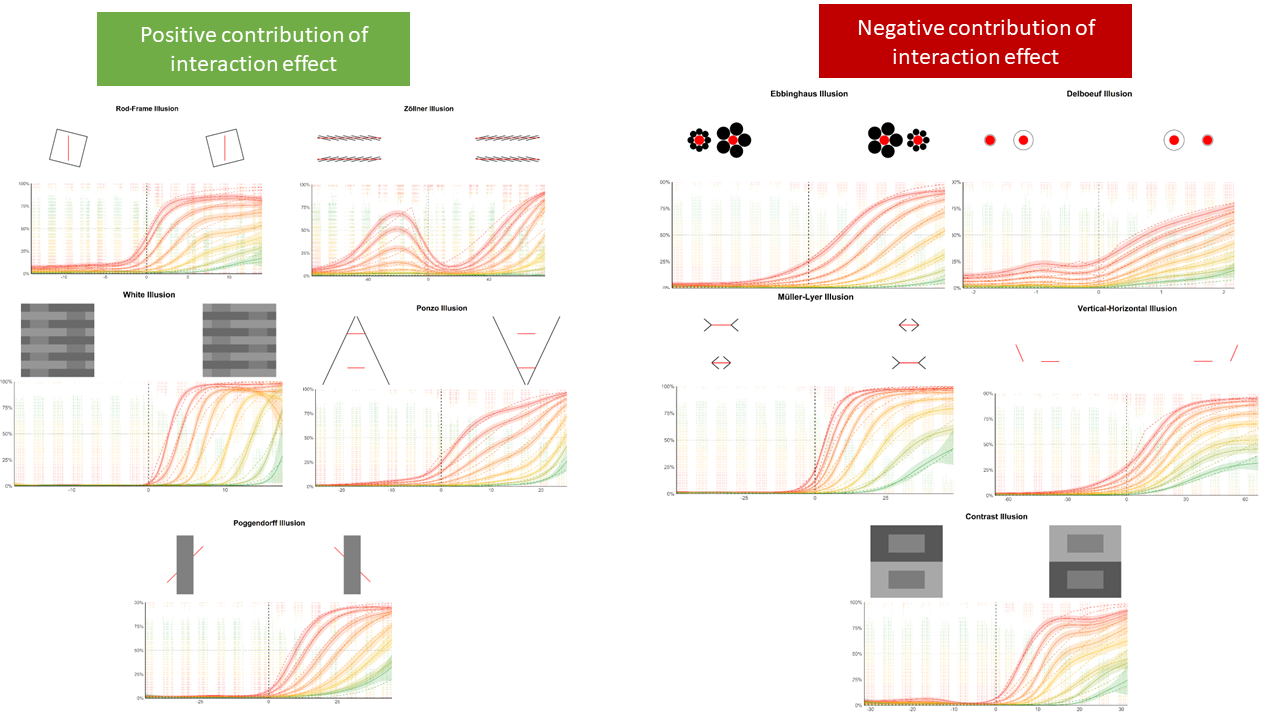
I particularly thank the reviewer for pointing that out, as this is indeed something that puzzled us for some time. We did not underline nor discuss this differential pattern in the manuscript to avoid too much speculations, but my hunch is that it has to do with the speed-accuracy trade-off and/or the over-power of some of the illusions.

To elaborate a bit, we modelled 3 regression parameters:

* Effect of task difficulty (length of the blue arrow): bigger negative numbers = when the task is more easy (at illusion strength of 0), the error rate / RT lowers more. This score was not included in the illusion-sensitivity scores as it is theoretically independent from it.
* Effect of illusion strength (slope of the green arrow): bigger positive numbers = when the illusion is stronger (at a task difficulty of 0), the outcome increases more. This is the slo
* Effect of interaction (the difference between the slope of the green arrow and the slope of the purple-ish arrow): this represents how much the effect of illusion (the slope of the green arrows changes as the difficulty decreases).



Now, from the plots alone, it is not obvious was distinguishes the illusions were the interaction effect has a negative contribution to the general score from those where it has a positive contribution:



We could argue that the group on the left has a more egg-shape like pattern, meaning that the effect for the easy difficulty (green lines) curves back up as illusion strength increases, closing up the space (particularly salient for White, Ponzo Poggendorff). Relatedly, it seems there is overall less interaction in the left group: the effect of illusion seems to be **shifted** (to the right), rather than **attenuated** (rather than of a decreasing slope). Good examples are White, where the slopes are similar, just starting at shifted locations of difficulty, vs. Muller-Lyer, where the pattern itself changes from an abrupt increase to a more progressive one.

However, even if true, the overall importance of an interaction (vs an independence of the two main parameters) between difficulty and strength does not alone explain why the interindividual variability in this interaction effect would load positively and negatively unto to the general factor. My hunch regarding the speed

Line 137 “After a brief demographic survey and a practice series of illusions” please clarify the phrasing used for each question.

TODO. Anshu.

How did the authors encourage/motivate over 1000 trials per participant in an online study? This in itself is impressive and it would be useful to report how long the task took, were participants allowed to complete over several sessions? Did all participants complete all trials? Did all participants complete all illusions?

TODO. Anshu.

Did all participants take part on laptops? Or was it possible on phones and tablets too It is mentioned that screen size was measured (Line 188) – how was screen size measured? Was a screen scaling method i.e. credit card scaling technique (Li et al 2020) included?

TODO. Anshu.

Figure 1 – It could be clearer what is meant by “direction” in these figures. I would suggest adding that to the description where task difficulty and strength are defined (top panel of figure 1 paragraph 2) . Perhaps placing boxes around the stimuli would also make it clearer which stimuli are paired together.

TODO. Dom.

Line 150 “The task was implemented using jsPsych” please can the authors also share how the study was hosted.

TODO. Anshu.

Line 156 “about £7.50” – please clarify what “about” means – was payment different for different participants?

TODO. Anshu.

Line 160 “implausibly fast” please define.

TODO. Anshu.

Figure 2 – For someone not familiar with each illusion it isn’t immediately clear how each image is “stronger” than the other. Perhaps also add to each what the question was i.e. “which is longer” to make it clear what each task entails.

TODO. Dom.

Figure 3 legend – please clarify what the x and y axes of the bottom plots correspond to i.e. they all look 0 centred?

TODO. Dom.

Figure 3 age distribution plot and personality train plots – need y labels.

TODO. Dom.