

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.

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 thank the reviewer for his positive comments and for the suggested papers. We have incorporated the relevant evidence from these papers accordingly (lines 76-81):

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 through the method of adjustment, which involves having them adjust the targets to perceptually match a reference stimulus¹⁶⁻³⁰. [...]

Lines 83-92:







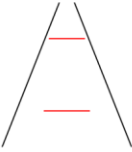



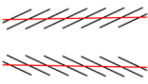









While some recent efforts have some made to implement more empirically rigorous paradigms³², most of the applied manipulations only focus on varying the physical dimensions of the illusion's target features without modulating its contextual elements, hence limiting the variability in the illusory effects of the stimuli presented. Furthermore, such prior studies have typically generated stimuli whose targets' physical attributes vary over a relatively narrow range, thus further constraining the reliability of their findings. As such, it is possible that the recent evidence reported against a common factor of illusions could be due to the low stimulus variance instead of a true reflection of a lack of common mechanism.

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.

We agree with the reviewer that it would be helpful for readers to have a more detailed description of the illusions used. We added a figure in the Methods section (see Figure 2):

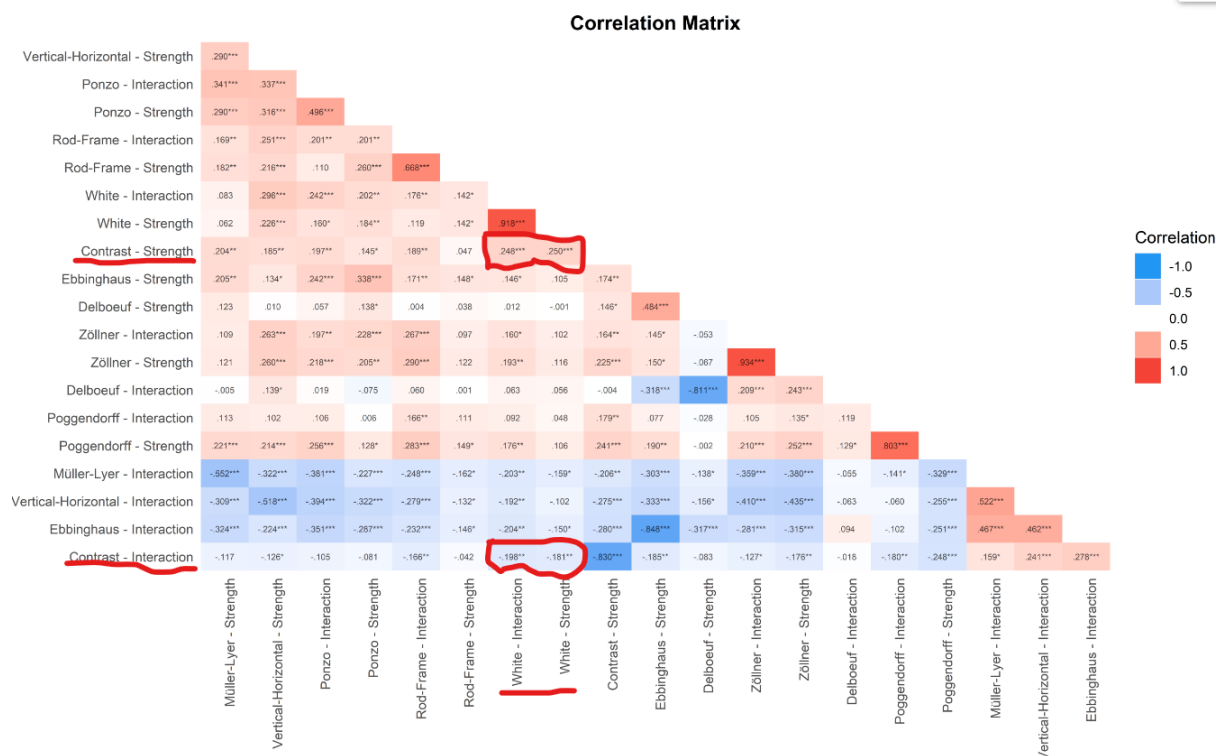
Illusion	Example	Task	Description
Delboeuf		Which red circle is bigger? 	Two circles surrounded by outer rings. The circle with the closer outline appears larger than the one with the distant outline.
Ebbinghaus		Which red circle is bigger? 	Two circles surrounded by other circles. The circle surrounded by smaller circles appears larger than the one with the larger surrounding circles.
Müller-Lyer		Which red line is longer? 	Two parallel segments that end with inwards/outwards pointing arrow-like fins. The segment with inward-pointing fins is typically perceived to be longer.
Ponzo Illusion		Which red line is longer? 	Two parallel segments embedded between a pair of converging lines. The segment on the convergence side is typically perceived to be longer.
Vertical-Horizontal		Which red line is longer? 	Two lines segments, one horizontal and one angled. The angled line is usually perceived as longer.
Zöllner		Which direction are the red lines converging towards? 	Represented by two long lines crossed with short, repeated lines, that appear as converging to one side.
Rod and Frame		Which direction is the red line leaning towards? 	A segment ("rod") enclosed in a tilted square ("frame"), that appears to be leaning to one side.
Poggendorff		Is the right line above or below the left line? 	Two oblique segments separated by a rectangle. The right segment appears to be vertically misplaced.
Simultaneous Contrast		Which small rectangle is lighter? 	Consists of two smaller grey rectangles embedded in backgrounds of different contrasts. The rectangle with the darker background appears lighter.
White		Which vertical rectangle is lighter? 	Composed of a series of contrasting horizontal bars superimposed with vertical grey rectangles. Rectangles superimposed onto darker bars appear lighter.

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 185-186):

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. Following the comments, 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. A few things can be noted: 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).



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

```
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.

We agree with the reviewer that the term ‘ambiguous percepts’ may be more well-suited to characterize a specific subset of illusions (e.g., bistable illusions) than as a general definition for visual illusions. We have since amended the terminology we use to define illusions accordingly (lines 63-65):

...illusions are conceptualized as non-veridical perceptual experiences that result from giving ample weight to prior knowledge to minimize prediction error in the face of biasing sensory evidence.

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.

We agree that this figure was a bit crowded. Given that we added a new figure in the methods, we thought that the best way to proceed was to remove the demographic information from the figure (as it was somewhat redundant with the sample description and did not add much relevant information), and to focus instead on the study results. Note that the figures related to the sample information and distributions is still available in the supplementary material.

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 to provide the most reliable data amongst other similar platforms (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 for experimenters to flag "bad" participants (a score that is also tracked by the platform which then removes consistently problematic participants from their platform), they are motivated to do the tasks thoroughly (also avoiding in 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:** Participants were informed of the task duration prior to the start of the experiment so that they 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 is a purely subjective feeling (and naturally biased), it seems to us that the nature of the stimuli makes for a fairly engaging experience (maybe not to the extent of "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 enhanced the impression of progression and allowed the participants to take short mental breaks when they desired to do 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 by creating a change.

Regarding outlier trials:

- **Block rejection:** we carefully inspected the reaction time distribution for each participant separately, for each block (see answer to reviewer 2 for an example). See figure https://github.com/RealityBending/IllusionGameValidation/blob/main/study2/figure_s/outliers_RT-1.png (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 scenarios (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 to doing the task after doing the personality questionnaires, they simply answered 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 RT 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

We started by clarifying that all regression models were fitted under the Bayesian framework (as it's more flexible and easier to model these different classes and families of models under that unified framework) (lines 185-186):

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 under a Bayesian framework.

We clarified the justifications (lines 187-192):

- GAMs:

We started by fitting General Additive Models (GAMs), which can parsimoniously accommodate possible non-linear effects and interactions.

- Logistic family:

Errors were analyzed using logistic mixed models (suited to estimate the error rate)

- ex-Gaussian family:

RTs of correct responses were analyzed using an ex-Gaussian family with the same fixed effects entered for the location μ (mean), scale σ (spread) and tail-dominance τ of the RT distribution^{38,39}

The reason to not only rely on GAMs can be found next (lines 193-195):

Using GAMs as the "ground-truth" models, we attempted at approximating them using general linear mixed models, which can be used to estimate the effects' participant-level variability (via random slopes).

As the goal was to extract participant-level scores, we needed random slopes for the effects, and the smooth terms of GAMs cannot be modelled as a random effect. Thus, we sought to additionally fit “simpler” models (i.e., general linear models with 2 parameters and their interaction), for which these parameters can be estimated for each participant (with all the benefits that it presents, in particular regularization of extreme scores).

EFA and SEM were used to give insights into the factor structure of the scores and test different possible hierarchical models, respectively.

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

Non-linear transformations were tested as non-linear relationships are typically observed in psychophysics and perceptual tasks. Linear, log, square-root and cubic-root offered a good coverage of the possible patterns (which indeed was the case as relative to the GAMs, at least for the illusions with monotonic effects). We clarified as follows (lines 195-199):

Following a comparison of models with a combination of transformations (raw, log, square root or cubic root; which are types of relationship commonly found in perceptual tasks) on the main predictors (task difficulty and illusion strength), we fitted the best model (based on their BIC and R2), and compared their output visually

As mentioned, we compared the matrix of combinations of transformations by comparing the BIC and R2 (based on these tables, found in the analysis script):

##	Name	BIC	R2_marginal	BF	Side	Model
## 1	DIFF--STRENGTH_sqrt	10894	0.7999		0.000772	Errors
## 2	DIFF_sqrt--STRENGTH_sqrt	10933	0.7828	< 0.001	0.000848	Errors
## 3	DIFF--STRENGTH_log	10952	0.8341	< 0.001	0.000857	Errors
## 4	DIFF_cbrt--STRENGTH_sqrt	10973	0.7786	< 0.001	0.000892	Errors
## 5	DIFF_sqrt--STRENGTH_log	10981	0.8136	< 0.001	0.000939	Errors
## 6	DIFF_log--STRENGTH	21404	0.0750		0.049386	RT
## 7	DIFF_log--STRENGTH_sqrt	21422	0.0740	< 0.001	0.051651	RT
## 8	DIFF_cbrt--STRENGTH	21429	0.0742	< 0.001	0.049425	RT
## 9	DIFF_sqrt--STRENGTH	21451	0.0735	< 0.001	0.049323	RT
## 10	DIFF_cbrt--STRENGTH_sqrt	21462	0.0727	< 0.001	0.051162	RT

We clarified as follows (lines 198-199):

we fitted the best model (based on their BIC and R2), and compared their output visually (Figure 3)

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.

We clarified as follows (lines 199-200):

Note that the model comparison and the parameters used in the resulting models were not pre-registered.

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?

EFA was used in an exploratory fashion to gain insights with a non-supervised method (and to see what happens when we “let the data speak”). SEM was then used to test models with layers based on empirical and theoretical hypotheses. We clarified as follows (lines 203-205):

using exploratory factor analysis (EFA, to gain insights into the structure), and structural equation modelling (SEM, to model and test different hierarchical models)

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

In the description of model parameters, all instances of “diff” refer to the task difficulty, which is characterized by the objective difference in the physical attributes of the targets (of

the behavioural task) within an illusion. This has been clarified in the Data Analyses section (lines 201-203):

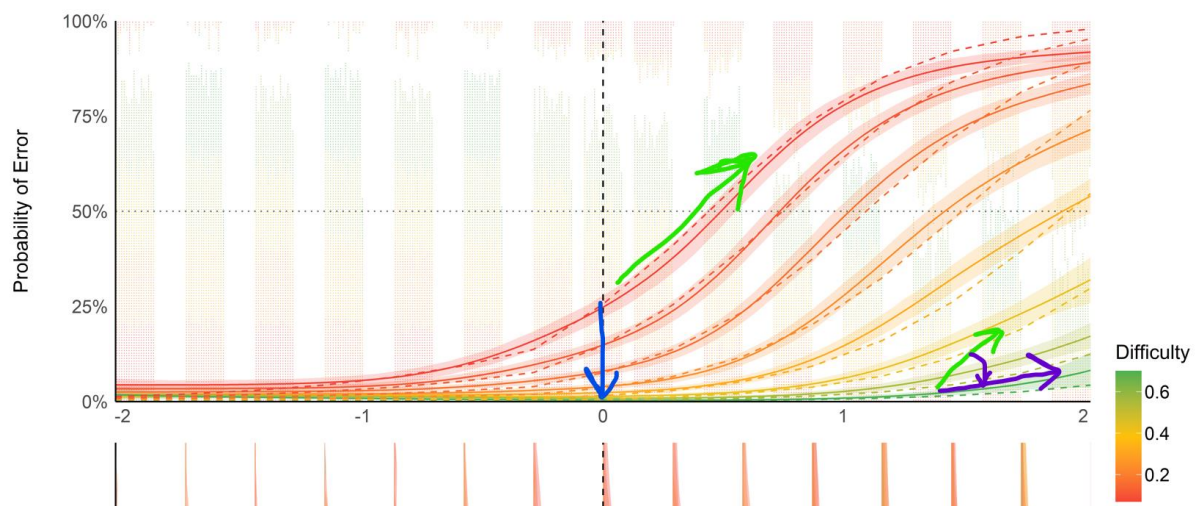
*The inter-individual variability in the effect of illusion strength and its interaction with task difficulty (**diff**) was extracted from the models and used as participant-level scores.*

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?

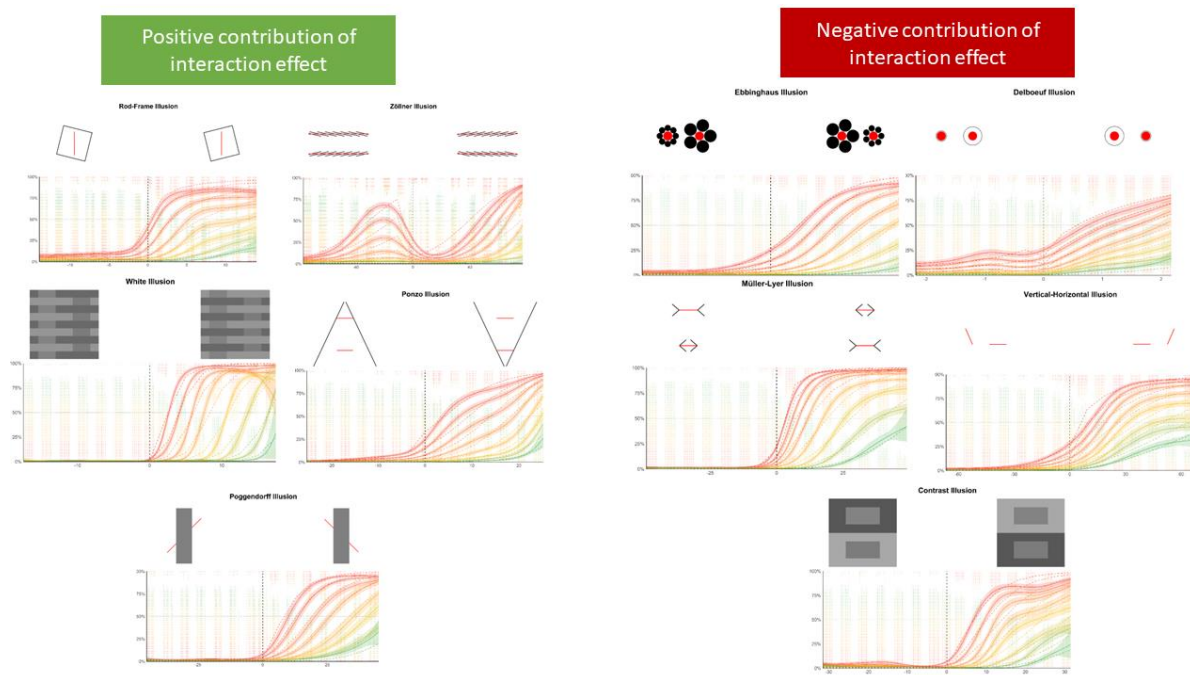
We particularly thank the reviewer for pointing that out, as this is indeed something that puzzled us for some time. We did not underline nor discuss this differential pattern in the manuscript to avoid too much speculation, but my hunch is that it has to do with the way participants sensitive to illusions respond to “easy” trials (low task difficulty).

To elaborate a bit, we modelled 3 regression parameters for each illusion:

- **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.
- **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 what distinguishes the illusions where the interaction effect has a negative contribution to the general score from those where it has a positive contribution (see groups below). Adding to the confusion, the difference in the sign of the interaction contribution does not seem to follow typical “groupings” of illusions (e.g., different signs for the 2 contrast-based illusions, or for the Ponzo and Müller-Lyer, even though these are typically considered relying on closer mechanisms).



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 with the red lines (this is 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** (i.e., a decreasing slope). The most striking example is 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 these observations were 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. What it seems to suggest is that participants with high illusion sensitivity have a weaker interaction effect for some illusions and a stronger one for others. My hunch is that it has to do with a change in the pattern of response between participants (and possibly the speed-accuracy trade-off strategy). Participants with a high sensitivity to illusions would be affected by illusion strength more progressively for easy difficulty trials (green lines), whereas we would observe a shift in the resistant participants (i.e., they would not be affected so much by the illusion in easy trials, up to a point, in which the effect of the illusion strength would follow the same pattern as for hard trials – but shifted).

At this stage, these thoughts are pure speculations. That said, this question is very interesting and would likely require a study on its own to replicate and elucidate.

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

We have clarified this in the Procedure section (lines 148-150):

Participants were first given a brief demographic survey, which collected information regarding their age, gender, country of birth, ethnicity, and highest attained education level.

The online task is available for readers to try at

<https://realitybending.github.io/IllusionGameValidation/study2/index.html> (the demographic part is at the beginning if one wants to see how it looked like). For reference, the exact phrasing of the items can be found in the experiment code (see <https://github.com/RealityBending/IllusionGameValidation/blob/79241e21c8f65c20271ede37aa7687320e886bde/study2/index.html#L140-L192>).

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?

We thank the reviewer for the feedback and suggestions. Indeed, various methods were employed to motivate participants. This includes gamifying the experiment (such as displaying scores following the completion of each illusion block) and presenting the stimuli in multiple phases to allow participants to take short mental breaks, hence attenuating effects of fatigue. (See above for our response to comment #5 by Reviewer 1 for a more detailed elaboration of the methods we applied to encourage participants across trials.)

With regards to the task's duration, the median time taken to complete the study was approximately 55 minutes. Participants were required to complete the task within a single session and were informed of the study's estimated duration prior to the start of the session to manage their expectations. We have added this accordingly to the Procedure section (lines 163-164):

All trials were required to be completed within a single session (total experiment duration: ~55 minutes).

In addition to the data processed and made available (including the outliers), 11 participants did not complete the experiment because they exceeded the time limit (120 minutes) - which was set by Prolific based on the estimated study duration – hence no data is available for them.

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?

The administration of the task was restricted to computers. This requirement was set on Prolific which then administers the experiment accordingly and prevents mobile phones & tablets from participating. Additionally, as the progression of the trials were dependent on responses via key presses not found on keypads on mobile devices ('left' and 'right arrow' keys), participants who had accessed the experiment using their mobile phones and tablets would not have been able to complete the task.

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

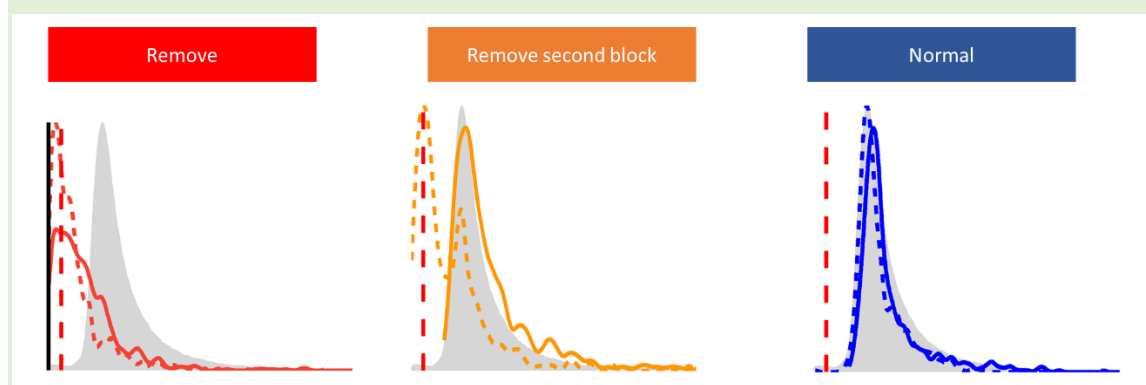
The payment rate (per hour) was fixed on Prolific, which then paid participants based on the median time taken to complete the experiment. As participant slots were opened in batches (to prevent overloading Pavlovia’s servers and facilitate quality control), there are minor fluctuations (of a few pennies) in the payment for each batch of participants caused by the variability of the median time. As such, we reported the average amount each participant was paid.

Line 160 “implausibly fast” please define.

To complete our answer to Reviewer 1 about ensuring data quality, “implausibly fast” refers here to 1) trials with less than 125ms (a lenient lower-bound threshold, much shorter than the P200 ERP wave before which meaningful perceptual processing + motor response is unlikely) and 2) participants with visually salient non-typical RT distributions, based on this figure:

https://github.com/RealityBending/IllusionGameValidation/blob/main/study2/figures/outliers_RT-1.png (takes time to load).

To illustrate, below are 3 participants with their RT distributions. The dashed-line correspond to the second block (after the personality questionnaires). The grey distribution area is the average distribution, which is quite narrow and well-defined (typical RT log-normal shape). The red dashed line is the 125 ms threshold for trial-wise exclusion. In blue is a typical participant, and in red is a participant which we labelled as outlier due to their salient pattern. In orange is one of the participants that has a first block that is normal (kept), and a bad second block (excluded). We think that these participants, after the personality questionnaires, when presented again with the same task as before, probably lost patience and tried to finish as fast as possible. Note that we also cross-checked with other indices, and indeed the error rate for our “outliers” was close to 50%, which would be the rate expected for random answers.



We clarified in the Participants section (lines 174-179):

We excluded 6 participants upon inspection of the average error rate (when close to 50%, suggesting random answers), and reaction time distribution (when implausibly fast relative to the average RT distribution). For the remaining participants, we discarded blocks with more than 50% of errors (2.16% of trials), possibly indicating that instructions were

misunderstood (e.g., participants focused on the shorter line instead of the longer one), and 0.76% trials with extreme response times (< 125 ms or > 4 SD above mean RT).

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.

We added the task question in the figure. Example for one illusion type:

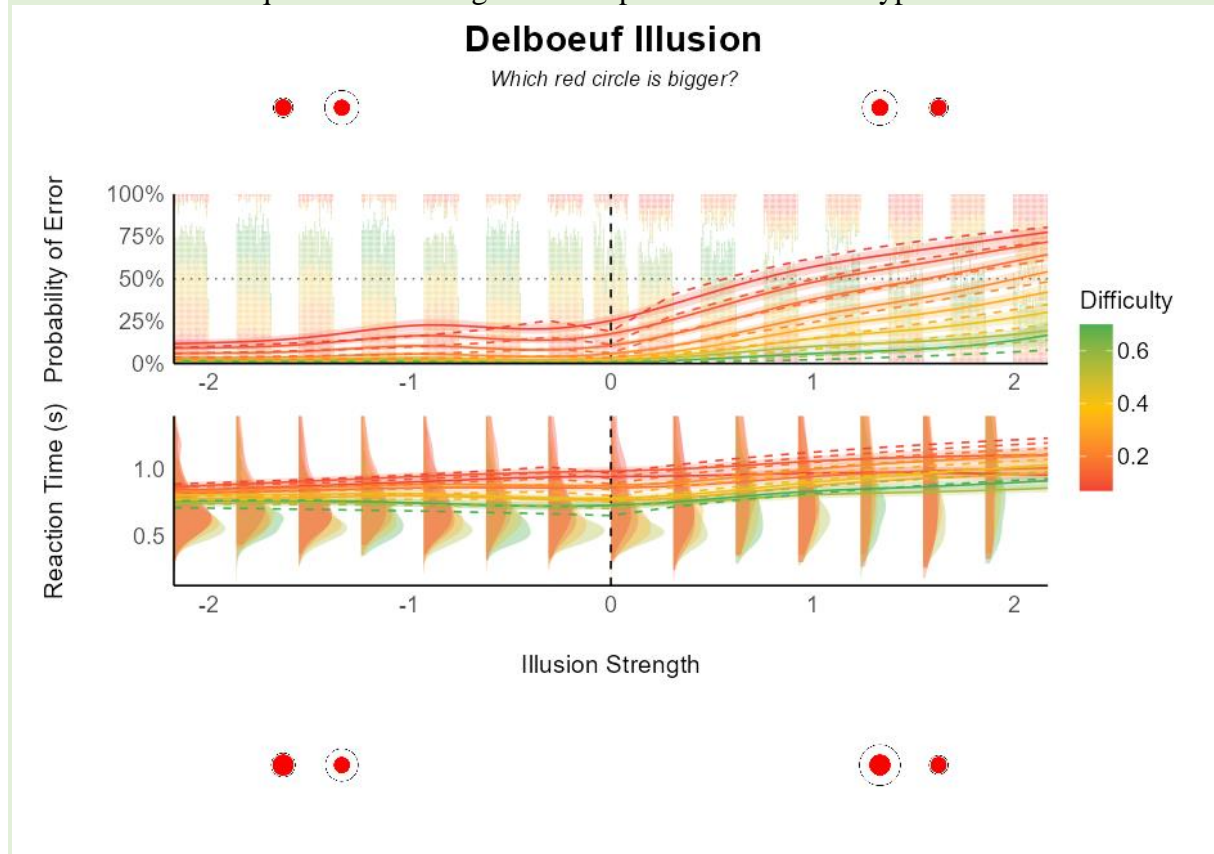


Figure 3 legend – please clarify what the x and y axes of the bottom plots correspond to i.e. they all look 0 centred?

We have clarified the legend as follows: (Note that this is now labelled Figure 4 following the addition of a new figure at the suggestion of Reviewer 1)

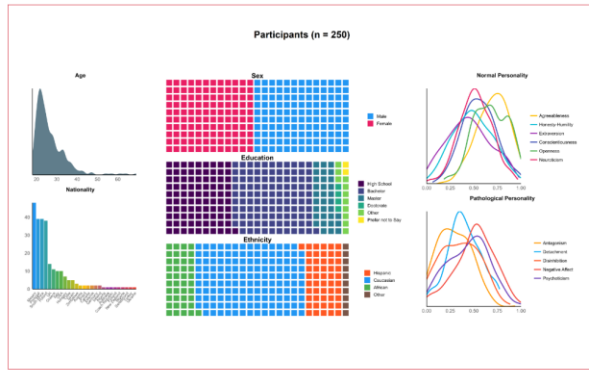
The upper plots show the illusion sensitivity scores as a function of sex and age (solid lines indicate significant relationships). Bottom plots show the correlation between the general factor (Factor i) of illusion sensitivity (on the x-axes) and personality traits.

Yes, the illusion scores are zero-centred (as they correspond to participant-level random effects, which are expressed in deviation respective to the “main” effect).

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

This has been added in the figure (which now is in the supplementary materials only, as we removed the demographic part from the figure following the suggestion of Reviewer 1).

Deleted



Added

