Appendix

Section A – List of critical items

Verbs ("is", "look"; "appear", "seem") were rotated across lists in a Latin Square Design. Asterisks indicate regions of interest.

1 *The dress*was*blue.*Hannah*thought it was*green.*
2 *The dress*was*blue.*Anna*thought it was*navy.*
3 *The cake's icing*seemed*rose-coloured.*Tim*believed it was*white.*
4 *The cake's icing*seemed*rose-coloured.*Tom*believed it was*pink.*
5 *The socks*looked*dark blue.*Michael*thought they were*black.*
6 *The socks*looked*dark blue.*Andrew*thought they were*indigo.*
7 *The colours*appeared*bright.*Greg*believed they were*dull.*
8 *The colours*appeared*bright.*Chris*believed they were*lively.*
9 *The hill*was*steep.*Peter*thought it was*gentle.*
10 *The hill*was*steep.*Paul*thought it was*sheer.*
11 *The courtyard*seemed*rectangular.*Phil*believed it was*square.*
12 *The courtyard*seemed*rectangular.*Bill*believed it was*oblong.*
13 *The medal*looked*elliptical.*Joe believed it was*round.*
14 *The medal*looked*elliptical.*Jim*believed it was*oval.*
15 *The chef*appeared*heavy.*Sam*believed he was*slim.*
16 *The chef*appeared*heavy.*Sue*believed he was*fat.*
17 *The cat darting across the street*was*small.*Ben*believed it was*large.*
18 *The cat carting across the street*was*small.*Matt*believed it was*little.*
19 *The building*seemed*quite grand.*Alex*thought it was*humble.*
20 *The building*seemed*quite grand.*Will*thought it was*stately.*
21 *The dog partially hidden by the fence*looked*massive.*Daniel thought it was*small.*
22 *The dog partially hidden by the fence*looked*massive.*William*thought it was*large.*
23 *The bathroom*appeared*reasonably large.*Julian*believed it was*small.*
24 *The bathroom*appeared*reasonably large.*Gabriel*believed he was*big.*
25 *The shop*was*shut.*Sophie*believed it was*open.*
26 *The shop*was*shut.*Ellie*believed it was*closed.*
27 *The fruit*seemed*overripe.*John*believed it was*fresh.*
28 *The fruit*seemed*overripe.*Ron*believed it was*rotting.*
29 *The jacket*looked*much used.*Edward*thought it was*new.*
30 *The jacket*looked*much used.*Gareth*thought it was*old.*
31 *The deer lying at the roadside*appeared*conscious.*Amy*believed it was*dead.*
32 *The deer lying at the roadside*appeared*conscious.*Kim*believed it was*alive.*
33 *The dog in the yard*was*gentle.*Grace*thought it was*dangerous.*
34 *The dog in the yard*was*gentle.*Sarah*thought it was*harmless.*
35 *The office*seemed*disorderly.*Dan*thought it was*tidy.*
36 *The office*seemed*disorderly.*Dick*thought it was*messy.*
37 *The word's spelling*looked*correct.*Olivia*believed it was*wrong.*
38 *The word's spelling*looked*correct.*Emily*believed it was*right.*
39 *The man*appeared*advanced in years.*Mary*thought he was*young.*

40 *The man*appeared*advanced in years.*Emma*thought he was*old.*
41 *The group's actions*were*organised.*James*thought they were*haphazard.*
42 *The group's actions*were*organised.*Joshua*thought they were*coordinated.*
43 *The statue*seemed*bronze.*Michael*thought it was*gold.*
44 *The statue*seemed*bronze.*Andrew*thought it was*brass.*
45 *The cutlery*looked*silver.*Jack*believed it was*steel.*
46 *The cutlery*looked*silver.*Joe*believed it was*sterling.*
47 *The flooring*appeared*wooden.*Sarah*thought it was*laminate.*
48 *The flooring*appeared*wooden.*Eve*thought it was*mahogany.*

Despite our norming efforts, two s-consistent items (numbers 10 and 14) and two s-inconsistent items (numbers 5 and 45) attracted mean plausibility ratings that were outliers (SD's >3) and were excluded from further analyses (i.e. plausibility and eye movements).

Section B – Further reading times

First Pass Reading Times

Verb Region. Results showed a significant effect of verb $F(3,138)=19.18$, $p<.001$, $\eta^2=.29$ (see Figure A). Paired comparisons revealed reading times were appreciably lower in 'is'-items than appearance verbs items (appear-is: $t(46)=5.88$, $p<.001$, look-is: $t(46)=-5.69$, $p<.001$, seem-is: $t(46)=-6.35$, $p<.001$). In contrast, the appearance items were not significantly different from one another (all *p's*>.18). The difference between 'is'-items and appearance items is easily explained by the shorter length and greater frequency of 'is' (Rayner, 1998).

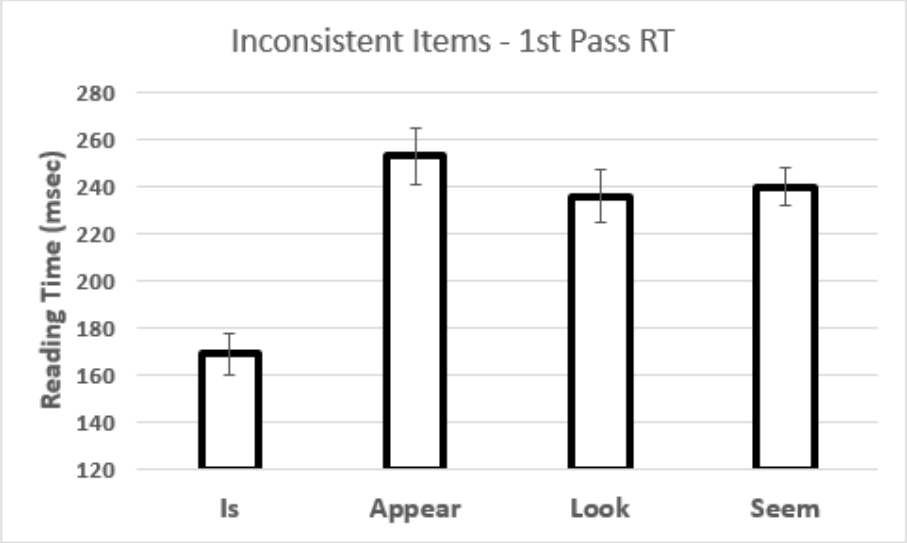

Figure A. Mean first pass reading times for the verb region in inconsistent items. Error bars show the standard error of the mean.

First Object. Results showed no main effect of verb $F(3,138)=.92$, $p=.43$, $\eta^2=.02$ (see Figure B). This means first pass reading times for the first object (e.g., 'blue' in item 1) were not significantly different for items with the contrast verb 'is' than for the appearance verbs.

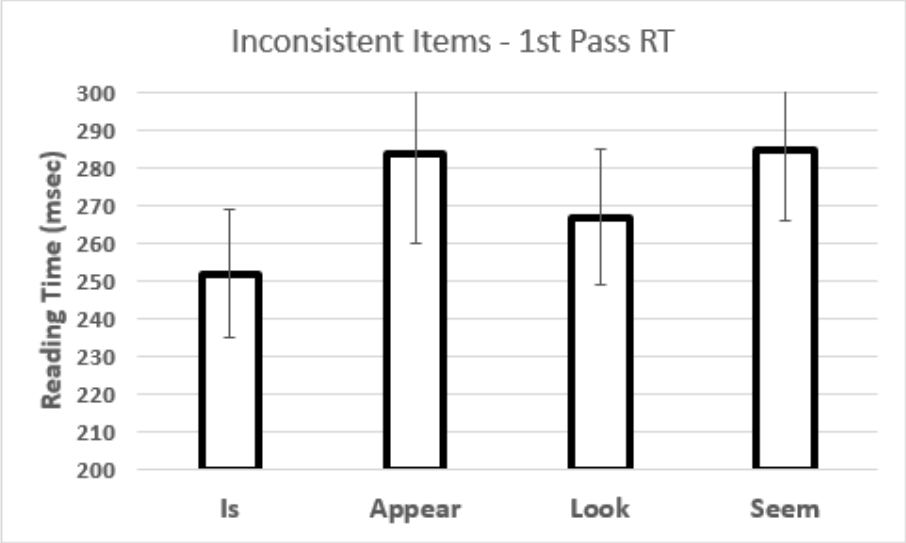

Figure B. Mean first pass reading times for the first object in inconsistent items. Error bars show the standard error of the mean.

Conflict Region. Again, there was no main effect of verb $F(3,138)=.75$, $p=.52$, $\eta^2=.02$ (see Figure C). This means first pass reading times for the conflict region (e.g., 'green' in item 1) were not significantly different for items with the contrast verb 'is' than for the appearance verbs.

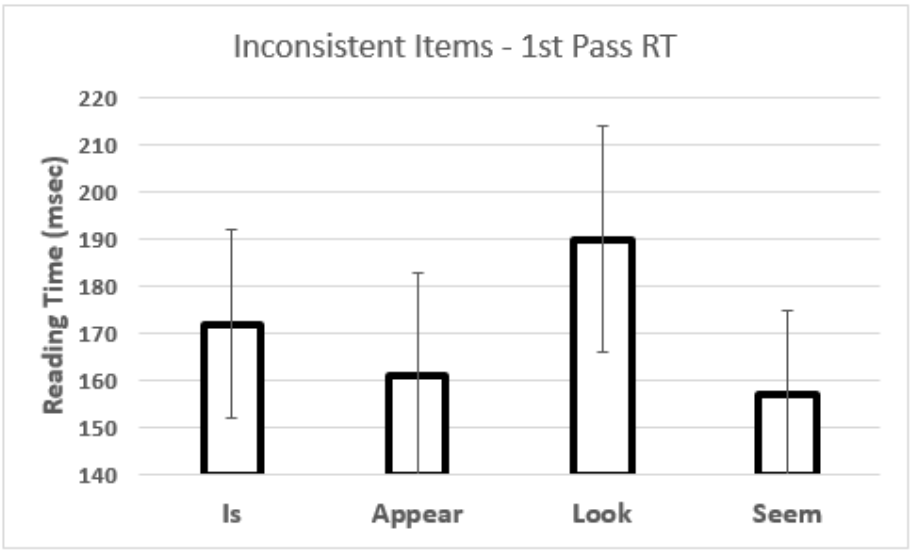

Figure C. Mean first pass reading times for the conflict region in inconsistent items. Error bars show the standard error of the mean.

Total Reading Time – Source Region

Total reading times on the source region (e.g., 'was blue', in item 1) showed a significant main effect of verb $F(3,138)=9.22$, $p<.001$, $\eta^2=.17$. The significant differences were between 'is' and the appearance verbs (appear-is: $t(46)=5.02$, $p<.001$, look-is: $t(46)=-2.57$, $p<.05$, seem-is: $t(46)=-4.48$, $p<.001$). The comparison between 'look' and 'appear' was also significant

$t(46)=2.11$, $p<.05$, the differences between 'appear'-'seem' and 'look'-'seem' were not *p's* >.08. Observed differences were driven not by first-pass reading times (above) but by second-pass reading times that are indicative of integration difficulties (Sec. 4.3).

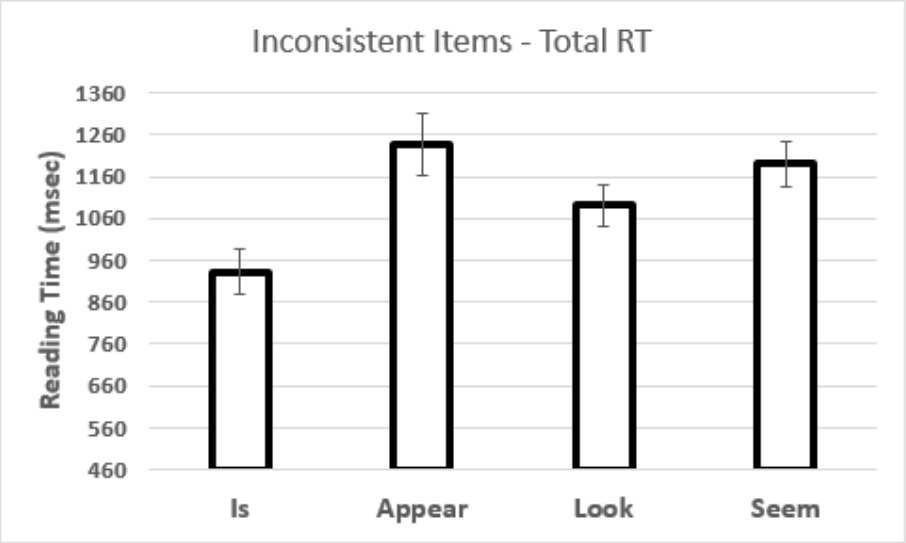

Figure C. Mean total reading times for the source region in inconsistent items. Error bars show the SE of the mean.

Second pass Reading Time – Conflict Region

Results showed no significant main effect of verb $F(3,138)=1.28$, $p=.28$, $\eta^2=.03$ (see Figure D). This is in line with first-pass reading times and the fact that while approximately 90% of trials showed a regression from the conflict region, only 22.2% of trials involved a return to the conflict region. This suggests that integration difficulties were addressed in re-reading the source region.

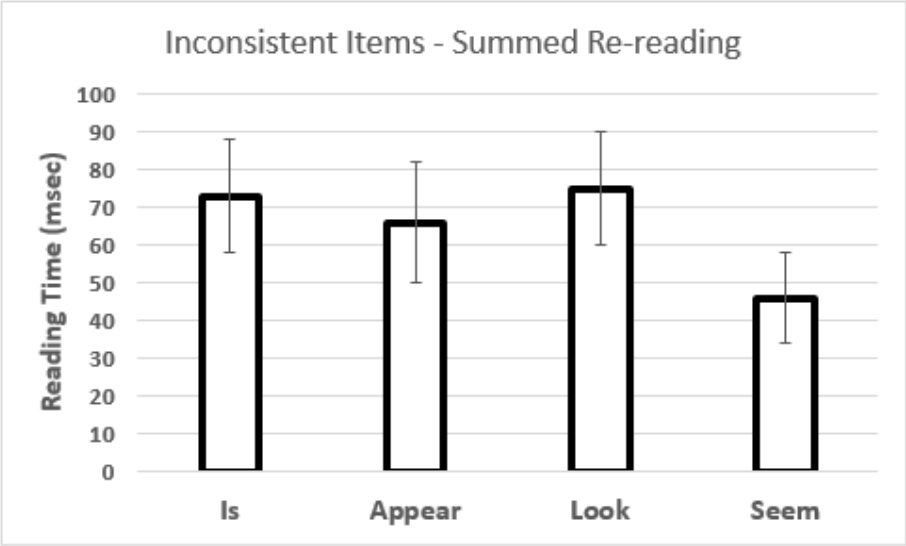

Figure D. Mean re-reading times for the conflict region in inconsistent items. Error bars show the SE of the mean.

Section C – Follow-up study on 'phenomenal' sense

We conducted an acceptability rating study to examine two hypotheses:

hI In ordinary discourse, all three appearance verbs have a 'phenomenal' sense or use that carries no doxastic implications (and allows us to say, e.g., that the round coin appears elliptical, when viewed sideways).

hII This sense or use is more salient for 'look' than for 'appear' and 'seem'.

Examination of hI is motivated by a reviewer query whether all three verbs have a phenomenal sense in ordinary discourse (or only in philosophy). Examination of hII is motivated by the mixed experimental picture concerning inappropriate doxastic inferences from 'look' (Sec. 4.5).

Methods

We recruited 51 participants with the same approach and from the same population as in Experiments 1-3 (Sec.3). They were all native English-speakers (74.5% women, three non-binary, with an average age of 36.5, ranging from 16 to 77). Participants received instructions that explained with the help of an example that the same word may be used in different senses, so that sentences that may appear to contradict each other can all be true, when using the word in different senses. They were told 'we are interested in whether the verbs "look", "appear", and "seem" are used in different senses that allow people to say different things that may appear to contradict each other'. They were instructed to 'take into account all senses of these verbs that you are familiar with', and asked to rate on a 7-point scale their confidence that given sentences using these verbs can be accepted as saying something true. Endpoints were explained as follows: '"1" means you are entirely confident that the sentence cannot be accepted as saying something true. "7" means you are entirely confident that the sentence can be accepted as saying something true. The mid-value "4" means you are really unsure whether or not the sentence can be accepted as saying something true.'

Participants rated 18 sentences that described familiar cases of non-veridical perception, where no adult perceiver is inclined to form a wrong belief about the shape, size, or colour of the object seen. Sentences used appearance-verbs either in a non-doxastic sense (to describe the looks of the object) or a doxastic sense (to state what property the viewer will think the object has), for example:

> When you look at a round coin sideways, the coin looks elliptical. (non-doxastic, ND)
> When you look at a round coin sideways, the coin looks round. (doxastic, D)

The first sentence is acceptable as true only if the verb has a non-doxastic reading (under the circumstances, nobody believes the coin is elliptical); the second is acceptable only if the verb has a purely doxastic reading (the viewer will believe the coin is round, but that is no description of what the coin 'phenomenally' looks like). Three pairs of sentences used shape adjectives, three used size, three used colour. Further examples include:

> Seen from the beach, the huge ships anchored out at sea appear small / huge. (ND) / (D)
> Under red light, white lab coats seem reddish / white. (ND) / (D)

Sentences were presented in random order. Verbs were rotated across items. Hypotheses concern non-doxastic (ND) items only. Doxastic items were intended as fillers, to ensure through contrast that participants would take into account the difference in adjectives (e.g., 'white' – 'reddish') in the critical ND items. We thus manipulated a single variable (verb) with

three levels. One participant was excluded as outlier (>3SDs from the mean in all conditions), prior to analysis.

Our analyses were driven by our *a priori* hypotheses. hI predicts that participants will accept ND items with all three verbs as saying something true (namely, in the non-doxastic sense), i.e., that acceptability ratings for ND items with all three verbs will be above neutral mid-point '4'. We assessed hI through three one-sample t-tests with a test value of 4. hII predicts higher ratings for ND items with 'look' than 'appear' or 'seem'. This results in a directional one-tailed hypothesis assessed via paired-samples t-tests.

Results and discussion

For ND items, mean ratings (with SDs) for 'look' (5.37, 1.09), 'appear' (5.03, 1.05), and 'seem' (4.97, 1.10) sentences were all significantly above mid-point 4 (appear: $t(49)=6.92$, $p<.001$; look: $t(49)=8.88$, $p<.001$; seem: $t(49)=6.23$, $p<.001$). Paired-samples t-tests revealed that non-doxastic 'look' sentences attracted ratings significantly different from counterparts with 'appear' $t(49)= -1.70$, $p=.048$ and 'seem' $t(49)=1.71$, $p=.047$. In contrast, ratings for ND sentences with 'appear' and 'seem' were not significantly different $t(49)= 0.31$, $p=.38$. For thoroughness, we also analysed results for doxastic items. Here, mean ratings (with SDs) for 'look' (3.12, 1.23), 'appear' (3.25, 1.25), and 'seem' (3.07, 1.09) sentences all were significantly below 4 (appear: $t(49)= -4.25$, $p<.001$; look: $t(49)= -5.05$, $p<.001$; seem: $t(49)= -6.00$, $p<.001$). Results for the paired-samples t-tests showed no significant differences between the conditions (all *p's* >.39).

Ratings for ND items were consistent with our hypotheses. However, since the doxastic sense or use of appearance verbs is well attested, in particular for 'appear' and 'seem' (Sec. 2.3), the ratings below neutral mid-point observed for doxastic (D) items with all three verbs suggest that participants answered a different question: How confident are you that the sentence is an acceptable (plausible or natural) thing to say about the situation envisaged? On this reading, the observed ratings for D items are consistent with the dominance of the doxastic sense at any rate for 'appear' and 'seem': Participants did not find it plausible or natural to use this – familiar – sense to talk about familiar situations of non-veridical perception. This may be because the circumstances indicated do not change judgment (adult perceivers are too conversant with them), but do change the way things look, so that participants inferred that the sentence must be about the way things look. On this reading, high ratings for ND items suggest participants found it plausible or natural to invoke a non-doxastic phenomenal sense to talk about the way things look. But this implies they recognised, and were familiar with, such a sense. Also on this reading, observed ND ratings therefore support hI. That participants find it more natural to use this non-doxastic sense of 'look' than 'appear' or 'seem' suggests the sense is more prototypical for 'look' (see Fn7). This is consistent with higher salience, as per hII.

Appearance verbs arguably are the linguistic devices best suited to talk about the way things look in cases of non-veridical perception. The fact that, even so, the proportion of '6/7'-ratings remained overall low for ND items (53%) suggests that participants frequently found it impossible to completely suppress doxastic inferences from these verbs, even with strong contextual support (stronger than in the main study). The fact that the proportion of such ratings was lower for ND items with 'appear' (47%) and 'seem' (51%) than 'look' (56%) suggests that participants found it easier to completely suppress doxastic inferences from 'look'. This is predicted by the Salience Bias Hypothesis in conjunction with hII and the corollary that the doxastic sense is less dominant for 'look' than the other verbs. We conclude that all three appearance verbs have a non-doxastic 'phenomenal' sense in ordinary discourse, and this sense

is more salient for ‘look’ than ‘appear’ and ‘seem’.

Section D – Discussion of H_2

The main study of the paper (reported in Sec. 4) also helps us assess the hypothesis H_2 as an alternative explanation of findings from a previous plausibility ranking experiment (Fischer et al., 2019). Instead of invoking inappropriate doxastic inferences, this explanation (Sec. 3.5) suggests that preferences for ‘is’ sentences in items like

The hill seemed / was quite steep. The rambler thought it was gentle

are based on how subjective vs objective the question under discussion (whether the hill is steep) is deemed to be: Participants prefer ‘is’ where this question is deemed objective, and appearance sentences where this is more subjective; the second sentences of items (‘The rambler thought it was gentle’) are not taken into account.

In the present main study, participants rated both items consistent and inconsistent with the doxastic inferences posited by H_1. This consistency manipulation in the second sentence affected plausibility ratings very strongly and more strongly than any other factor examined. This suggests that plausibility assessments also in the prior plausibility ranking study were influenced by the second sentences that affected consistency. Furthermore, in the present study, the consistency manipulation affected the plausibility of sentences with different verbs (slightly) differently, as we observed a (marginal) verb × consistency interaction. This suggests that where verbs have stronger stereotypical associations with doxastic patient properties, doxastic inferences reduced the plausibility of s-inconsistent sequels more strongly. We therefore infer that, in the earlier study, preferences were influenced by differences in verbs’ strength of stereotypical association with doxastic inferences that were cancelled by the second sentence.

Crucially, the follow-up study (Sec. 4.4) showed that higher subjectivity ratings either failed to change plausibility ratings (in the s-consistent condition) or changed the plausibility of both appearance and ‘is’-sentences in the same direction, and to pretty much the same extent (in the s-inconsistent condition). Differences in subjectivity ratings therefore cannot explain the observed differences in plausibility ratings between items with ‘is’ and with appearance verbs ‘appear’ and ‘seem’, in the main study, or the preferences observed in the earlier plausibility ranking study by Fischer et al. (2019) (with only s-inconsistent items).

Finally, H_2’s alternative account requires that the patient role of verbs in appearance sentences be assigned to the author, rather than the protagonist, of the text (Sec. 3.1). This predicts a positive correlation between subjectivity and plausibility ratings for appearance sentences (Sec. 4.1). The findings of the follow-up study excludes this confound at any rate for items with visual objects. (The present study examined no others.) These findings speak against H_2 and suggest that, in the earlier study by Fischer et al. (2019), at any rate the preferences for ‘is’-sentences over appearance sentences in items with visual objects were based on making and maintaining contextually inappropriate doxastic inferences about the protagonist.

Fischer and colleagues (2019) further observed attenuated preferences in items with abstract objects. They adduced this to a moderate amount of patient reassignment in such items, due to higher levels of perceived contradictoriness (Sec. 2.3). This explanation is inconsistent with the fact that, in our new Exp.1, s-inconsistent appearance items with visual and with abstract objects were deemed equally contradictory, at any rate in items with verbs ‘seem’ and ‘appear’ (Sec. 3.2). The negative correlation between subjectivity and plausibility ratings observed in the follow-up study for s-inconsistent items with visual objects (Sec. 4.4) provide

an alternative explanation for patient reassignment in items with abstract objects: Without patient reassignment, participants infer that protagonists have beliefs or opinions that clash with those attributed to them by the sequel; protagonists entertain clashing doxastic attitudes. The negative correlation means that participants find it less plausible that a protagonist has clashing doxastic attitudes towards matters of subjective opinion than towards matters of objective fact. This might reflect epistemic doubt: Participants take s-inconsistent items to convey that protagonists are second-guessing themselves. Second-guessing oneself may intuitively make more sense when the matter at hand is objective rather than subjective.

In Exp.3, most items with abstract objects were perceived as more subjective than most items with visual objects: Breaking down the 36 items by quartiles based on descending subjectivity rating, we observe: Of the nine most subjective items, all but one item had an abstract object (88.9%). Of the second nine items, all but two had an abstract object (77.7%). By contrast, of the least subjective nine items, all but one had visual objects and of the second to bottom nine all but two had visual objects. This means that 15 of 18 items with an abstract object (83.3%) were in the more subjective half of items, while 15 of 18 items with visual objects (83.3%) were in the less subjective half.

Even though appearance items with visual and abstract objects are perceived as equally contradictory, items with abstract objects tend to be perceived as more subjective, so that second-guessing oneself intuitively makes less sense and items are perceived as less plausible. Principles of charity have readers take remedial action to avoid gross implausibility. This extra source of implausibility may therefore push some appearance items with abstract objects over the threshold at which some participants resort to patient reassignment to obtain a more plausible interpretation of appearance sentences. This could account for the attenuation in preferences for ‘is’-sentences observed by Fischer and colleagues (2019). We therefore take present findings to reinstate the findings from that previous study in the light of a potential confound.