Grimes and Higher-Order Theory

David Rosenthal
CUNY Graduate Center
Philosophy and Cognitive Science
https://www.davidrosenthal.org/

June 10
2022

ACCELERATING RESEARCH ON CONSCIOUSNESS

TEMPLETON WORLD CHARITY FOUNDATION

I. Higher-Order Theory

- Higher-order (HO) theories all endorse in some form what I have called the transitivity principle (TP):

  A necessary condition for a mental state to be conscious is that one is in some way aware of being in that state. There is a higher-order awareness (HOA) of the state.
- That HOA is seldom conscious; we're rarely subjectively aware of it. And though a HOA may rest on some inference, no such inference is ever conscious. So we have a subjective sense that our HOA of the first-order (FO) state is direct and transparent.

- It's plain that if one is in some mental state but is in no way aware of being in it, that state is not conscious.
  And that's logically equivalent to TP.
- If one is aware in a suitable way of being in a state, there is something it's like to be in that state—there is subjective awareness.
- There are variant HO theories. Mine posits a conceptual HOA—a higher-order thought. On most variants the HOA is a state distinct from the FO state it makes one aware of. But we can in any case always distinguish HO content from FO mental properties. So HO theories are always two-factor theories.
  But all that matters here is TP, and I'll usually just speak generically of HO theory.

- I am not here arguing for HO theory—just clarifying what it does, and doesn't, say.
- Because HO content is distinct from FO mental properties, it's conceivable that the HOA might misrepresent the FO properties. One would then be aware of oneself as being in a state that one is not actually in.
  But HO theory does not imply that such HO misrepresentation ever happens. The theory is neutral—it simply leaves it open.
- Most theories don't distinguish these two factors, and so can't accommodate such misrepresentation. So their advocates claim any acceptable theory must preclude it—though if it doesn't ever occur one could just add that stipulation to HO theory.

II. HO Misrepresentation

- Because HO theory is neutral, and makes no prediction about whether subjective awareness does ever misrepresent FO states, it's no argument against HO theory if such misrepresentation never occurs.
- But if HO misrepresentation does occur—however rarely—that supports HO theory.
- Because the HOA is distinct from FO mental properties, HO theory can accommodate misrepresentation by subjective awareness, and can readily explain it if it does occur.

  And any theory that has the relevant two-factor structure needed to accommodate misrepresentation is perforce a HO theory.

- No other theory allows a disparity between subjective awareness and FO mental properties. (Maybe GWT: Pre-change FO content is available, but not post-change.)

  The relevant misrepresentation occurs only if subjective awareness is wrong about the type of FO mental state an individual is in.
- Fixation is irrelevant to determining what the HOA is. The HOA is not a visual state, but a subjective awareness of what visual state one is in. And that's determined by subjective report, not visual functioning.
- HO misrepresentation is not one's being wrong about what one sees.
  It's when subjective awareness is wrong about what one's actual visual state is.

III. Our Findings

- This is all to clarify what HO theory says—and doesn't say—about misrepresentation. Let's then turn to our findings.
- The main finding for HO theory is that after missing a change, subjects about half the time report the changed item as having its pre-change attribute.
- And we can assume that the visual state—the state in visual cortex—did change. If so, the FO state represents the post-change attribute and the HOA represents the pre-change visual state. Subjective awareness then misrepresents the FO visual state.

- That's the prima facie situation. Let's consider and assess some concerns.
- Perhaps the FO state didn't always change. But all that's needed is that it changed some of the time. Just a single case of subjective misrepresentation of a visual state will imply HO theory, since (almost) no other theory can explain even that.
- Subjects on miss trials divided 50-50 in reporting pre- vs. post-change attributes. So were they just guessing? No. A 50-50 split shows nothing about individual cases.

  And if they were guessing, it would be about their subjective memory of what they last saw, since they're no longer seeing the item. So these are subjective reports.

- Slides 9-13 added June 24:
- We want to know what the post-change subjective awareness was on miss trials. Foil questions about pre- vs. post-change attributes were designed to reveal that.
- The guessing issue is delicate. If replies were random, they'd plainly be unhelpful about post-change subjective awareness.
- But guessing need not be—and often isn't—random. Subjects can be unsure but still take their best guess. Taking one's best guess without being sure is not random; it's what one takes to be so—just without complete certainty. Confidence ratings could identify any guesses that were truly random—leaving nonrandom best guesses.

Grimes and Higher-Order Theory June 24, 2022 9

- Foil questions occur after the post-change display has vanished. So responses rely on subjective memory—one's best sense about the items most recently seen. And that reflects one's last relevant subjective awareness, which was post-change.
- So probing for post-change subjective awareness does not require a high level of confidence—just enough to rule out random guessing.

  And moderate confidence—really anything more than minimal—will do that, and so will reveal post-change subjective awareness.
- And again: A few cases of post-change subjective awareness of the pre-change attribute are enough to support HO theory.

Grimes and Higher-Order Theory June 24, 2022 10

- Adding 'don't know' or 'opt out' as replies to foil questions would create problems. One couldn't tell whether those replies meant subjects just weren't totally sure—as against their really having no idea at all, so that it was genuinely random.
- And that distinction is what's pivotal for guessing and related typles of response.
- Reaction times (RTs) are a good surrogate for confidence ratings. We found that RTs were statistically the same across response types—and tended not to be long enough to suggest random guessing.

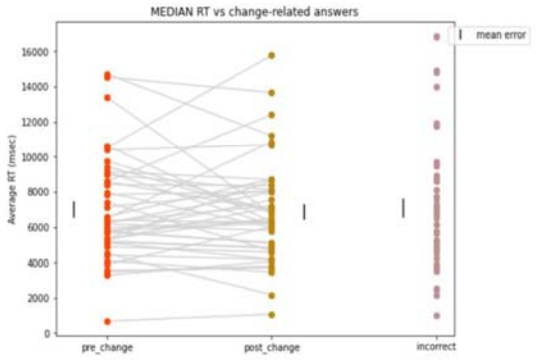
  This doesn't show there was no random guessing—just that it was not prevalent. And as just noted, that's all that's needed.

- Summary of RT findings:

Response time by response given to the change-related Q

For each participant, we took the median RT across trials for each response (pre-change, post-change, and incorrect).

One participant's RT was exceptionally short (the one at the bottom in the figure). We removed him/her in the statistical analysis.



- Because subjects plainly see post-change displays consciously, there is always some subjective awareness to find out about.
- There are two possibilities on miss trials. (1) The post-change subjective awareness was of the post-change attribute—but the subject nonetheless registered no change. Or (2) post-change subjective awareness was of the pre-change attribute—and that explains why no change was registered.
- Even if some subjects were random on some responses to the foil questions, one of those two things was the case. We need the best evidence about which of those possibilities obtained. It's a mistake to ask for absolute certainty.

Grimes and Higher-Order Theory June 24, 2022 13

- One might still question whether subjects' responses to post-change queries reflects their post-change subjective awareness. Thus a concern was raised that subjects' responses might just reflect a memory of the pre-change attribute without thereby expressing a subjective awareness of it.
- But if their subjective awareness did change when the FO visual state changed, subjects wouldn't respond to post-change queries from their pre-change memory. And if their subjective awareness did not change, it did misrepresent a changed FO state.
- Also, nonrandom responses to post-change queries are the best evidence we can have about their subjective awareness.

Grimes and Higher-Order Theory June 10, 2022 14

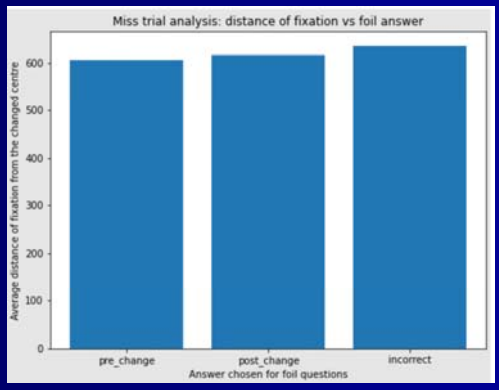
- There is no test for subjective awareness more basic than subjective report.
  Any so-called no-report test must itself be validated by appeal to subjective report.
  This is not at all special to HO theory; it's accepted in all consciousness research.
- Most subjects who reported the pre-change attribute also didn't re-fixate to the changed item—a striking finding, to which I'll come back in just a moment.
- But again, that's not in any case relevant to what the HOA is, which is determined solely by subjective report—in our work, by responses to the post-change queries.
  Indeed, those queries are arguably an important advance in our replication.

IV. The FO States

- Though fixation is not relevant to what the HOAs are, it could make a difference to whether the FO visual states changed.
  If fixation remains far from the item that changed, the stimulus change might not cause any change in the FO visual state.
- The post-change FO visual state would then just reflect the pre-change attribute.
  A post-change HOA that reflected the pre-change attribute would then correctly represent the post-change FO visual state.
  And without HO misrepresentation of the FO state, there's no support for HO theory.

- But all that is likely only if fixation remains far from the changed item.
  If fixation is not that far from the changed item, the FO visual state will likely change to reflect the change in the stimulus.
  And distant fixation is the only reason why the FO visual state might not change.
- Also: FO visual states definitely do change with reports of a post-change attribute; if they didn't, subjects would not be aware of that post-change attribute.
  So if fixation is alike for reports of pre- and post-change attributes, FO states almost certainly also change even when subjects report pre-change attributes.
- And our results here are striking.

- Distance of fixation from the changed item on miss trials does not on average differ significantly between reports of pre- and post-change attributes—indeed, it's even a tiny bit closer for pre-change reports:



- So FO states almost certainly do change with reports of pre-change attributes.

- Other work, say with fMRI, will further nail down changes in FO visual states. But our results already strongly support FO visual changes without changes in subjectivity.
- Subjective awareness also diverges from FO visual states in unconscious change detection (Thornton & Fernandez-Duque 2001), and in long-lasting postdictive effects (Michel & Doerig 2021), such as chromatic flicker fusion (Jiang et al 2007). So we should not resist evidence of misrepresentation here—
and consequent support of HO theory.

Thanks for your attention!

REFERENCES

Thornton, Ian M., and Diego Fernandez-Duque (2001), "An Implicit Measure of Undetected Change," Spatial Vision, 14, 1 (January): 21–44.

Jiang, Yi, Ke Zhou, and Sheng He (2007), "Human Visual Cortex Responds to Invisible Chromatic Flicker," Nature Neuroscience 10, 5 (May): 657-662

Michel, Matthias, and Adrien Doerig (2021), "A New Empirical Challenge for Local Theories of Consciousness," Mind & Language, online March 17, doi: 10.1111/mila.1231