Loss of Confidence statement

2025-07-22

Around 2017, Christopher Chabris, Tal Yarkoni, Julia Rohrer and others conducted the Loss of Confidence Project, in which authors were encouraged to publicly declare instances where they were no longer confident in the claims made in one of their prior publications. [Rohrer et al. (2021)](https://journals.sagepub.com/doi/10.1177/1745691620964106) is absolutely worth a read.

I saw it at the time when they were recruiting co-authors, but one of the requirements was that participants had to take full responsibility for the work they had lost confidence in (in order to avoid finger pointing at other coauthors of the original work, etc.). The work I had already lost confidence in at the time was conducted for my bachelors thesis, and I felt at the time that I would unfairly take full responsibility for analytic decisions that my supervisor instructed me to take. In retrospect, this was splitting hairs, and I regret not being braver, getting involved in the project and being radically transparent.

Over the years, I’ve often considered posting a loss of confidence statement like this. The more (forensic) meta-science work I do, the more important it feels to be transparent about what of my work I (currently) am confident in or not. I recently wrote a separate piece on [my recent and current research interests](https://mmmdata.io/research/). This reignited thoughts about a loss of confidence statement, which is effectively the other side of the coin: the publications which I would encourage prospective students and colleagues not to read. So, here it is.

**tl;dr: As with much of the psychology literature, disregard a lot of the empirical work from before about 2015.**

### Hussey & Barnes-Holmes (2012)

Hussey & Barnes-Holmes (2012) The Implicit Relational Assessment Procedure as a measure of implicit depression and the role of psychological flexibility. \*Cognitive and Behavioral Practice. <https://doi.org/10.1016/j.cbpra.2012.03.002>

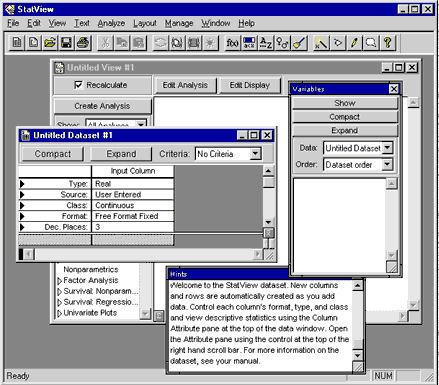
**Summary**

* Confidence in computational reproducibility: ✅ (probably)
* Confidence in replicability: ❌
* Confidence in measurement validity: ❌
* Confidence in claims: ❌
* I recommend you do not cite, attempt to build on, or even read this article.

**Details**

This work was conducted for my bachelors thesis project in 2009. It’s *p*-hacked to death, on the instruction of my supervisor - although that word didn’t even exist until 2010, and I probably didn’t learn it until 2012 or really appreciate its meaning until 2015 or so. Pick your poison, they’re all in there: optional stopping, flexibility in scoring, median split of a continuous variable, mulch-way interaction effects.

My recollection is that I successfully computationally reproduced the reported results from the final data set around 2015, when I first started learning R. The original analyses were conducted in statview, a statistics package not even sold since 1997, but which somehow refused to die in my supervisors research group - it continued to see use there until at least 2015. Unfortunately, I don’t have the records for this any more, and the point is moot given that the reported analyses are just one of many, many forking paths.



I’ve subsequently written lots (e.g., see [this post](https://mmmdata.io/posts/2025/04/irapresearch.org-website-relaunched/)) on my more general loss of confidence in the validity of the IRAP, the measure used in the study.

More importantly, I’m aware of at least one unpublished independent conceptual replication of the study that failed to replicate the group differences or interaction.

I have told people about all of the above privately for years, but it’s worth being explicit here: I have lost confidence in the claims contained in that article, including but not limited to the claim that depression is associated with differences in the patterns of implicit bias / brief and immediate relational responding related to depressive statements.

The current norms of science are such that it is practically unheard of to self-retract an article “merely” for *p*-hacking, so I have not made an attempt to do so. When I’ve asked colleagues for their opinions on this over the years, the feedback I get is that it would be akin to self-flagellation. Nonetheless, I continue to be concerned by it in a way I can’t shake, given that it continues to be cited (116 citations on Google scholar at time of writing).

### O’Regan, Farina, Hussey, & Roche (2015)

O’Regan, Farina, Hussey, & Roche (2015) Event-related brain potentials reveal correlates of the transformation of stimulus functions through derived relations in healthy humans. Brain Research. <https://doi.org/10.1016/j.brainres.2014.12.044>

**Summary**

* Confidence in computational reproducibility: 🤷‍♂️
* Confidence in replicability: 🤷‍♂️
* Confidence in measurement validity: ❌/🤷‍♂️
* Confidence in claims: 🤷‍♂️

**Details**

This work was conducted very early in my PhD using a small grant we obtained for between-research-group collaborations. It is Lego Science: I had a behavioral task ‘block’ (the IRAP) and they had a neuro ‘block’ (EEG), neither of us deeply understood the other’s, but we clicked the blocks together to build a publishable unit. It’s not so much that I have lost confidence in it, rather I’m not sure I ever really had it. Not for any deep concern, just a general sense of bewilderment.

### Finn, Barnes-Holmes, Hussey, & Graddy (2017)

Finn, Barnes-Holmes, Hussey, & Graddy (2017) Exploring the behavioral dynamics of the implicit relational assessment procedure: The impact of three types of introductory rules

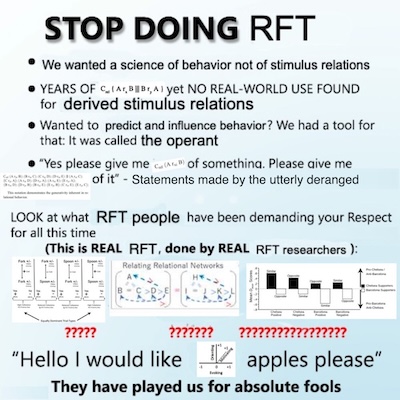
**Summary**

* Confidence in computational reproducibility: ✅
* Confidence in replicability: ✅
* Confidence in measurement validity: ❌
* Confidence in claims: ❌

**Details**

I’m confident that the general finding is replicable, i.e., that IRAP trial types that should intuitively show equivalent effects do not, because it has been both directly and conceptually replicated by at least three teams (see [Hussey & Drake, 2020](https://osf.io/preprints/psyarxiv/sp6jx/)). However, the specific claims regarding the explanations for this effect presented in this article are, in my opinion, vague just-so stories based on weak evidence and questionable analyses.

The claims in article related to what the authors call the STTDE continued the [Gish Gallop](https://en.wikipedia.org/wiki/Gish_gallop) present in a lot of the Relational Frame Theory literature, and specifically Dermot’s work, whereby new acronyms for theories and concepts were created with much more enthusiasm that they were ever tested, from AARR, BIRR/EERR, REC, DAARE, STDDE, MDML, HDML, and so forth. That theoretical car crash is beyond the scope of this post, and likely better to [satirize](https://bsky.app/profile/ianhussey.mmmdata.io/post/3lrfiourvvc2y) than critique given the claims are [Not Even Wrong](https://en.wikipedia.org/wiki/Not_even_wrong):



The authorship process for this article is also worth commenting on. These studies were conducted over a few years, beginning with one I designed and ran in Maynooth and then later with ones run by the other authors in Gent. The write up process happened in 2016, after was no longer working for Dermot Barnes-Holmes. At the time, I had been raising concerns about the IRAP as a measure and the analysis practices in his research group for a handful of years, and the professional relationship was already quite strained. As far as I recall, the manuscript for this article was given to me on a Thursday afternoon, and I was told that I had a one hour meeting the next day to give any feedback I had on it, at which point it would be submitted. The manuscript was, therefore, a fait accompli. With more experience, I would have put my foot down and asserted my right and responsibility as an author to stand over all claims in the manuscript.

After publication, I requested the data for the studies I had not been involved in to attempt to reproduce the published results. Strangely, the other authors were quite resistant to sharing the data with me, despite me being a co-author. I remember posting to the psychMAP facebook group - an important marketplace of ideas in the early days of the Credibility Revolution - to ask for help and advice with the situation. Brian Nosek replied with encouragement that denying co-authors access to data they have to put their name to is not tenable, which gave me the confidence to push harder and eventually obtain it.

The results did computationally reproduce, as far as I can recall. However, the appropriateness of some analyses and the conclusions drawn are inappropriate. For example, continuous variables are dichotomous (oof) and then subjected to a Chi-square test (double oof, this is just a less powerful t test), and then labelled an “individual level analysis” in the article (triple oof, it is still very much a group level analysis). This point was not merely statistical pedantry, it participated in several intractable back-and-forth arguments in peer reviews in subsequent years, where authors and reviewers argued about whether or not the IRAP was claimed to have individual level utility, e.g., based on whether this article really meant what it stated or not. These debates were a case of Schrodinger’s claim: it allowed proponents of the IRAP to say that they did vs. did not ‘really’ claim that the IRAP had individual level utility, depending on the context.