**Was this Registered Report pilot tested? Examination of Vaidis, Sleegers, Van Leeuwen, DeMarree, Sætrevik, Ross, ... & Priolo, D. (2024)**

Vaidis, D. C., Sleegers, W. W., Van Leeuwen, F., DeMarree, K. G., Sætrevik, B., Ross, R. M., ... & Priolo, D. (2024), published in AMPPS, included 104 authors from 39 research teams attempting to replicate the induced-compliance paradigm in study 1 Croyle and Cooper (1983). The study was preregistered and receives public marks for transparency and rigor. Yet, one thing that is transparently missing is any mention of pilot or pretesting the study to make sure it works before having the 39 labs try to replicate with those materials. This leaves the entire effort open to small problems, since it is also the first time most of the teams were trying a study like this. Thus, we ask what piloting mechanisms if any were employed, and suggest to the community more generally that it makes sense to make sure a protocol is working before sending it out to e.g., 39 places.

Words: 1,362

Vaidis et al., (2024) brought together 39 teams to try and replicate the induced-compliance paradigm in study 1 Croyle and Cooper (1983). The study is reported as a failure with the authors reporting that, ‘The primary analyses failed to support the core hypothesis’... and that ‘Overall, the results call into question whether the induced-compliance paradigm provides robust evidence for cognitive dissonance’ (abstract). While the paper scores highly for transparency, the paper or registration documents do not report any pre or pilot testing of the study before having e.g., 39 teams try it, which is curious and deserves more attention.

The basic argument is that it makes sense to pilot test and make sure the study is really working before trying it across 39 teams. This will also help improve the replicability rate of the field. If the project was ‘properly’ pilot tested, it should have worked at least once and this should be transparently reported. Conversely, if the goal is to show that the study does not work, having 39 teams try for their first time using an unpiloted protocol is also not the best way.

Single failed replication projects that do not include pilot testing or multiple attempts are less ideal than ‘positive’ replications which are able to ‘certify’ that something does or will work. Such large effort and expensive replication attempts should probably be reserved for showing what most likely does work rather than what might not work, especially if it will be most research teams first time trying a study in the area.

**If the materials were piloted, it should be reported that it worked at least one time**

The project has a significant focus on transparency, which makes it surprising that they do not report trying to make sure it works before sending it out to the 39 teams. If the materials were pilot tested it should be reported to have worked at least once, or if the materials did not work, they could have changed it so that it does work.

Piloting the work once to make sure everything is running smoothly before trying it across e.g., 39 sites simply makes sense. Registered Reports work to establish a new level of transparency and quality in the mind sciences, and this should include high quality piloting

protocols, and these should be reported in the transparent (registered) report. These tests will prevent such large scale failures to replicate, and help the field have a better 'replicability rate'.

**Does it make sense to try it in 39 places if we cannot get it to work in a single place?**

If the effect cannot be produced in a single place, then it is probably not worth examining whether it works across 39 sites. Still, we hope the community agrees that testing it at a single site should be a prerequisite for testing it across 39 sites. The way that the study was designed suggests that the authors believed that the study would work but forgot to test it before sending it out. This suggests a small methodological error, rather than that the entire field is not true.

These questions are not idle, and there have been convincing examples of small errors that set a whole series of replications on the wrong track. One of the best known examples is Wagenmakers et al., (2015) and the resulting Many Smiles Collaboration (Coles, 2019). Wagenmakers led 17 labs in trying to replicate the facial feedback hypothesis (Strack, Martin, & Stepper, 1988). For transparency and ease they decided to videotape participants, but this was later shown to be important in determining whether the effect appeared or not (Noah, Schul, Mayo, 2018). Finally the Many Labs effect confirmed the existence of a small but real facial feedback mechanism existing, despite the failure to replicate from Wakenmakers et al., (2015).

Especially when the 39 studies fail, the fact that it was not piloted becomes a significant problem for the research team, because they (should) then have to explain why it did not work and why they did not pilot it. It is literally designing a process without having done it well before.

**Piloting matters even in and especially for registered reports**

Some may argue that they should not need to pilot a study if they do not make substantial changes to the original report (which could in theory act as a pilot). This is not a strong argument because there are many things where even with very clear instructions, many people cannot do it, for instance putting furniture together or making fire with sticks. We additionally know that there are important details about the methods which are not actually and

explicitly mentioned in the methods section of published papers (Brenninkmeijer, Derksen, & Rietzschel, 2019).

Especially if we will treat the results of the study as particularly definitive, the study should be done to the highest standards, and this probably includes some pilot testing to make sure everything is working before the big effort replication. Unfortunately, and this is not the only instance of it (Wagenmakers et al., 2015), this study appears to have been done without having been piloted, perhaps in part leading to its failure.

The question is whether this registered report is really testing the replicability of the effect, or the replicability of the effect on one's first unpiloted attempt.

**Who gets better results, 39 people doing it the first time or one person doing it 39 times?**

Another way to think about this study is e.g., making shoes. This project brought together 39 teams who had never made e.g., a particular pair of shoes before (i.e., forced compliance paradigm), even if they make belts, purses, or even other shoes in their normal jobs - they all went to school together and have similar degrees.

Notably, these 39 teams came together to replicate the best results produced by a team that was working on the project for some years. That is, their first, even non-piloted, efforts are being compared to the best that the original researchers can produce, those they chose to send to high impact journals. In addition to it being the replication team's first time trying to make this pair of shoes/ effect, the set of instructions and materials for making the effect had never actually been used before being sent out, meaning that the instructions could be flawed.

The question is who will get 'better' results, the team that tried some times and submits their best results or the group of 39 that each tried once. We believe the team that was able to do it a few times and learn between the trials will produce better and more true results. In the end it is an empirical question, but it is notable that the Many Labs project have asked people to try a single time and report on their results. The results might be different if teams are able to try a few times while learning before reporting their 'best estimate' results.

**Using big effort replications to indicate what really works rather than what might not**

Given their expense, Registered Replications are probably better used as a ‘gold standard’ for a field in terms of registering that something is going to work across many sites and circumstances. This would mean that the study and set of materials should work in e.g., a high school classroom and could thus be used to teach the field. This would naturally include some pilot testing to ensure that everything is working smoothly before sending it out to the big time test and hopefully successful report. Such will also produce a better ‘reproducibility’ rate.

**Conclusion - Especially large scale studies should be piloted before their ‘best estimate’**

Large scale replication efforts are setting new standards for quality and excellence but should still be pilot tested, especially before the big/ best effort replication attempt across e.g., 39 teams. Not only should piloting prevent many of the major failures from happening, thus improving the replicability rate of the field, it will give the researchers a basis from which to start investigating why it failed, if it does (since we know it worked at least once).

**References**

Brenninkmeijer, J., Derksen, M., & Rietzschel, E. (2019). Informal laboratory practices in psychology. *Collabra: Psychology, 5*(1), 45.

Coles, N. A., March, D. S., Marmolejo-Ramos, F., Larsen, J. T., Arinze, N. C., Ndukaihe, I. L., ... & Liuzza, M. T. (2022). A multi-lab test of the facial feedback hypothesis by the Many Smiles Collaboration. *Nature Human Behaviour*, *6*(12), 1731-1742.

Croyle, R. T., & Cooper, J. (1983). Dissonance arousal: physiological evidence. *Journal of personality and social psychology*, *45*(4), 782.

Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, *114*(5), 657.

Strack, F. (2016). Reflection on the smiling registered replication report. *Perspectives on Psychological Science*, *11*(6), 929-930.

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A non obtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768 –777. http://dx.doi.org/10.1037/0022-3514.54.5.768

Vaidis, D. C., Sleegers, W. W., Van Leeuwen, F., DeMarree, K. G., Sætrevik, B., Ross, R. M., ... & Priolo, D. (2024). A multilab replication of the induced-compliance paradigm of cognitive dissonance. *Advances in Methods and Practices in Psychological Science*, *7*(1), 25152459231213375.

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Zwaan, R. A. (2016). Registered replication report: strack, martin, & stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917-928.