Possible Interpretations of a Failure to Replicate
Open Science Collaboration
Material adapted from Open Science Collaboration (2012) chapter about the Reproducibility Project

A possible outcome of the Reproducibility Project is that one, some, or many of the original studies will not replicate. Aside from the non-trivial question of determining whether the results indicate replication success or failure, what does it mean if a study fails to replicate? The simple, and inaccurate, answer would be that the original effect is therefore false. That could be the reason, but there are additional considerations. A failure to replicate could indicate that:

1. **The original effect was false.**

The original result could have occurred by chance (e.g., setting alpha = .05 anticipates a 5% false-positive rate), by fraud, or unintentionally by exploiting researcher degrees-of-freedom in design, analysis, or reporting (Greenwald, 1975; John, Prelec, & Lowenstein, 2012; Simmons, Nelson, and Simonsohn, 2011).

2. **The replication was not sufficiently powered to detect the true effect (i.e., the replication is false).**

Just as positive results occur by chance when there is no result to detect (alpha = .05), negative results occur by chance when there is a result to detect (beta or power). Most studies are very underpowered (see Cohen, 1962, 1992). Fair replications attempts should not follow that trend. Setting power at .95 means that the false-positive and false-negative rates should be equivalent for the anticipated effect size. With replications, an informed decision can be made for the anticipated effect size - the effect size of the original demonstration. If the original demonstration significantly overestimated the actual effect size, then a high-powered test may still miss a true (but much smaller) effect.

3. **The replication methodology differed from the original methodology on features that were critical for obtaining the true effect.**

There is no such thing as an exact replication. A replication necessarily differs in sample (i.e., even if the same participants are used, their state and experience differs) and setting (i.e., even if the same location, procedures and apparatus are used, the history and social context has changed). There are infinite dimensions of sample, setting, procedure, materials and instrumentation that could be conditions for obtaining an effect. However, effects are not interpreted as existing only for the original circumstances and having no informational or explanatory power outside of that lone occasion.

Part of standard research practice is to understand the conditions necessary to elicit an effect. Does it depend on the color of the walls? The hardness of the pencils used? The demographics of the sample? The social context of measurement? How the materials are administered? There are an infinite number of possible conditions, and a smaller number of plausible conditions, that could be necessary for obtaining an effect.

Plausibility is defined by intuition, reason, or accumulated knowledge. A replication attempt will necessarily differ in many ways from the original demonstration. The key question is whether its design is plausibly changing critical conditions to obtain the effect. There are three types of plausibility violations to consider:

1. *Published constraints on the effect*. Does the original interpretation of the effect impose conditions on the effect that are violated by the replication attempt?[1] If the original interpretation is that the effect will only occur for women, and the replication attempt includes men, then it is not a fair replication. The existing interpretation (and perhaps empirical evidence) already imposes that constraint. Replication is not expected.

2. *Constraints on the effect, identified a priori*. When reviewing the design of the replication, can design features be identified that might disrupt replication, even if they had not been described in prior published work. Published scientific reports are abridged summaries of the actual research. It is not possible to verbally describe all of the features of an experimental context - both because some are not linguistically translatable and because a complete specification would require infinite space to describe all of the conditions. Researchers do their best to provide an abridged summary of what they believe to be the critical features of the study design. But, these are necessarily imperfect. When observing an attempt to replicate, unpublished design features that are understood to be important, may be identified. In this sense, *a priori* means known (or presumed) constraints that are not part of the published record. Formally, it does mean that the published record is incorrect, but the information is still known or presumed as part of the tacit knowledge of a person, lab, group of labs, or even an entire field.

3. *Constraints on the effect, identified post hoc*. Constraints identified beforehand are distinct from the reasoning or speculation that occurs after a failed replication attempt. There may be many real, unknown constraints on the reproducibility of an effect, but they can only be discovered by conducting replications and varying the circumstances of data collection. That means that the replication attempt is fair, because the constraints could only become known by conducting the replication. This is the ordinary process of discovery. An effect turns out differently than expectation, so an information search ensues to figure out why. Researchers identify new plausible reasons. Most of these

[1] "Original" is inclusive of all prior published work on the effect if there are already multiple published replications.

reasons will presume the validity of both findings - original and replication - and seek to identify an explanation for the difference between them. The identification of a plausible explanation is the beginning, not the end, of empirical verification for the difference between the results. Because there are infinite differences between any replication and its original, this process can continue indefinitely. It could lead in at least two directions: (1) so much constraint on the original result that it is more sensible to conclude that it is false, or (2) constraints on the original result that affirm its truth but provide new, more nuanced, understanding of its meaning. That is, the original effect is not reproducible as originally interpreted, but is reproducible with the newly discovered constraints.

4. Errors in implementation, analysis, interpretation for the original, replication or both

Errors happen. What researchers think they did, or report doing, might not be what they actually did. Discrepancies in results can occur because of mistakes. There is no obvious difference between “original” or “replication” studies on the likelihood of errors occurring.

The Reproducibility Project considers all of these in its evaluation of replication attempts, and reporting on the results. Some can be addressed effectively with design. For example, all studies will have at least .80 power to detect the original effect, and the power of the test will be evaluated as a predictor for likelihood of replication. Also, differences between original and replication methods will be minimized by obtaining original materials whenever possible and by collaborating with original authors to identify and resolve all possible published or *a priori* identifiable design constraints. Finally, original authors and other members of the collaborative team will review and evaluate the methodology and analysis to minimize the likelihood of errors in the replications.