What are possible reasons for a non-replication:

Broad strokes:

- **there could be issues in either the original or replication**
- **in any of sampling, methods, or analysis**

Some of these are "fixable" and some will cause non-replicability

Note that non-replications could be multiply-determined.

A non-exhaustive chart of reasons for replication failure:

	Sampling	**Methods**	**Analysis**
Original	- Result is very specific to the original population - Small sample + publication bias led to an overestimate of the effect size	- some unspecified details are crucial to the result ("hidden moderators")	- Bug in original code creates a false positive - p-hacking / exploiting analytic flexibility
Replication	- Noisier data or treatment less effective due to platform mismatch (in person vs. online) - Cultural or temporal factors were different from original - Replication was underpowered to detect original effect size	- Stimuli weren't updated to account for cultural/temporal difference (ex. political) -procedure changes made the task unclear/too hard -manipulation/induction fails - some change to procedure is actually crucial to the effect	- Bug in replication code (or data recording)

What differences were there between the original and first replication? (A lot of these were switching from in-person to online, what are the potential consequences of that?)

- in person vs. online
- Participants paid less attention to some instructions
- Possibly different stimuli
- Same stimuli, but different implementations (maybe the color of the buttons or something)
- Replication was run in a different time (might be salient for questions about, e.g., the 2008 election)

Which of these problems can you resolve and what's your plan for them?

- Code problems: you should review the code in the original replication
- Original replication was underpowered (for any number of reasons) -> run a bigger sample
- Manipulation / induction didn't work -> possibly try to create a stronger induction (enforce reading passage or writing on topic better?)
- Stimuli are culturally/temporally specific -> update to appropriate time period/culture (carefully)

Which of these problems can be diagnosed and what's your plan for that?

- Power problems: run a power analysis for the smallest effect size of interest / a smaller effect size (then if negative result, you can basically rule out a large effect)
- Manipulation / induction didn't work (and there was no manipulation check) -> add a manipulation check
- Participants were careless / Data is noisy -> add (appropriate) attention checks and time how long participants are taking.

Your goal should be to be able to say which causes you can or can't rule out and what size effect you are able to detect.