The Replication Crisis Explained

N. WAGNER & R. TERNES

MI AIR

NOV. 2018

Replication as the Cornerstone of Science

- We all know that the scientific method relies on careful observation, prediction & hypothesis formation, testing, and then analysis.
- But replication and reproducibility is a critical element of the scientific method as well.
  - Without it, we cannot be sure that science was even done.
  - In this way, it serves as the cornerstone of the entire scientific process – ensuring the veracity of claims made by its practitioners.

Half of all published findings on human subjects are wrong.

- That probably includes the research that you've done too.

- This is a bold claim, is there any evidence to support this?
  - Yes, and quite a lot too. A few citations are included in the next slide.
  - Ionnadis (2005) suggests that less than half of medical findings are true
  - Open Science Collaboration replicated 100 studies in psychology, and less than half of the effect sizes in the replications were in the 95% CI of the original study.
  - A 2018 Study (Nosek et al.) found that only slightly more than half of studies in Nature and Science replicated, with effects sizes on average half of what was reported. These studies were pre-registered!

A Few Citations…

- Ioannidis JPA (2005) Why Most Published Research Findings Are False. PLoS Med 2(8): e124. https://doi.org/10.1371/journal.pmed.0020124
- Colhoun HM, McKeigue PM, Davey Smith G (2003) Problems of reporting genetic associations with complex outcomes. Lancet 361: 865–872.
- Gelman, A. & Loken, E. The statistical crisis in science. *American Scientist* **102**, 40 (2014).
- Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. Doi: 10.1126/science.aac4716
- Nosek et al. (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior, 2018.* DOI: 10.1038/s41562-018-0399-z
- Simmons, Nelson, & Simonsohn (2011) False-Positive Psychology: Undisclosed Flexibility in Data Collection & Analysis Allows Presenting Anything as Significant. *Psychological Science*. 11, 1359-1366. https://doi.org/10.1177/0956797611417632

ESP Broke the Camel's Back

- Methodological problems within psychology and other fields had been brewing for quite some time.
- But, in 2011 Daryl Bem published a paper in the Journal of Personality and Social Psychology that provided 10 'well-controlled' studied that demonstrated ESP.
  - Participants literally predicted the future. Or so the article claims.
- Statisticians and methodologists skewered the article.
  - All known replication attempts have failed.
  - However, no one believes that Bem created fake data.
  - Instead, he essentially p-hacked his way to statistical significance.
  - Because of the outlandish claims of the article, the resulting firestorm ignited a more serious investigation into the methodological shortcomings of current statistical practices within the field of psychology and beyond.

P-Hacking 101

- P-Hacking is a method of statistical inquiry that ensures an experimental result reaches statistical significance, even though your theory is total bunk.
- Here's how to do it:
  - Test lots of hypotheses.
  - Examine results as they come in. Stop collecting data when the results reach statistical significance.
  - Small samples sizes can help boost effect sizes. Especially when used in conjunction with the first two recommendations.
  - Choose a broad and vague theory that can make predictions in any direction. Then test in only one direction.
  - Definitely HARK (set up your Hypothesis After the Results are Known)
  - Find a way to throw out data that rejects your theory. (Outliers, or 'warm-up' trials).
  - Find ways to create subgroups, especially if they don't make sense or are irrelevant to your theory. (Age, gender, race/ethnicity, SES, height, weight, only child, etc.)
  - Include control variables in your study – but do so after looking at the results (feel free to remove control variables until your model matches your conclusion).

Type I & Type II Errors

- Another reason why replication and reproducibility is important
- False Positive (When the null is true but is rejected – as false by the testing). – Type I
- False Negative (When the null is false but is accepted). – Type II
- Happens in Medicine & Social Sciences – Could be consequences!

Reality	*Null (H_0) not rejected*	*Null (H_0) rejected*
Null (H_0) is true.	Correct conclusion.	**Type 1 error**
Null (H_0) is false.	**Type 2 error**	Correct conclusion.

Null	***Type 1 Error: H_0 true, but rejected***	***Type 2 Error: H_0 false, but not rejected***
Medicine A does not relieve Condition B.	Medicine A does not relieve Condition B, but is not eliminated as a treatment option.	Medicine A relieves Condition B, but is eliminated as a treatment option.
Consequences	Patients with Condition B who receive Medicine A get no relief. They may experience worsening condition and/or side effects, up to and including death. Litigation possible.	A viable treatment remains unavailable to patients with Condition B. Development costs are lost. Profit potential is eliminated.

Type M and Type S Errors

- Many of you probably recognized the classic 'Type I and Type II' errors.
- But there are other errors that researchers have identified recently.
- Type 'M' errors are errors about the magnitude of the effect size,
  - Occurs more often when small studies find significant results and no one bothers to replicate them to show that the effect size is smaller than reported.
- Type 'S' errors are errors about the sign (positive or negative) of the effect.
  - Just as worrisome as Type 'M' errors!
  - Has similar causes.

How to Reduce False Positives

1. Try to make direct replications of your own work if you can.
2. Use hold out samples.
3. Interrogate your results – are they sensitive to changes in methodology? If yes, then reduce your internal confidence that the results represent a real pattern.
4. Commit to sample size estimates before hand. Write it down, stick to it.
5. In general, stop doing one-sided statistical tests.
6. Don't even run tests on low Ns. Just don't.
7. Be wary of large effect sizes. Extraordinary claims require extraordinary evidence.
8. Exploratory analyses are great! But remember that a p-value does not have the same meaning in exploratory research as it does in confirmatory research.
9. Strong theories are good. Weak or vague theories are indistinguishable from exploratory analyses.
10. Multilevel modeling can help too, especially if you have lots of comparisons.

The impact on the quality of science

- Obviously, the items presented on the previous slide are not commonly practiced? Right?
  - Actually, some of them are standard practice in a number of fields and subfields.
    - Several of the articles in the citation slide have wonderful examples of systemic issues in medicine, neuroscience, genetics, and psychology.
  - Many researchers without extensive backgrounds in statistics may seriously underestimate how much some of these issues can increase false positive rates.
- In addition, it is shockingly easy for subtle biases to creep in – even for experienced researchers and statisticians.
  - Wanting to be right is part of human nature.

Replication to the Rescue

- Replication isn't necessarily a cure-all.
- But, it can help tremendously.
  - Direct replications are far more impactful than conceptual replications.
  - But conceptual replications can still help triangulate understanding.
  - They have a way of keeping us honest with ourselves too.
  - Also serve as a way to reveal fraud.

Conceptions of Replication

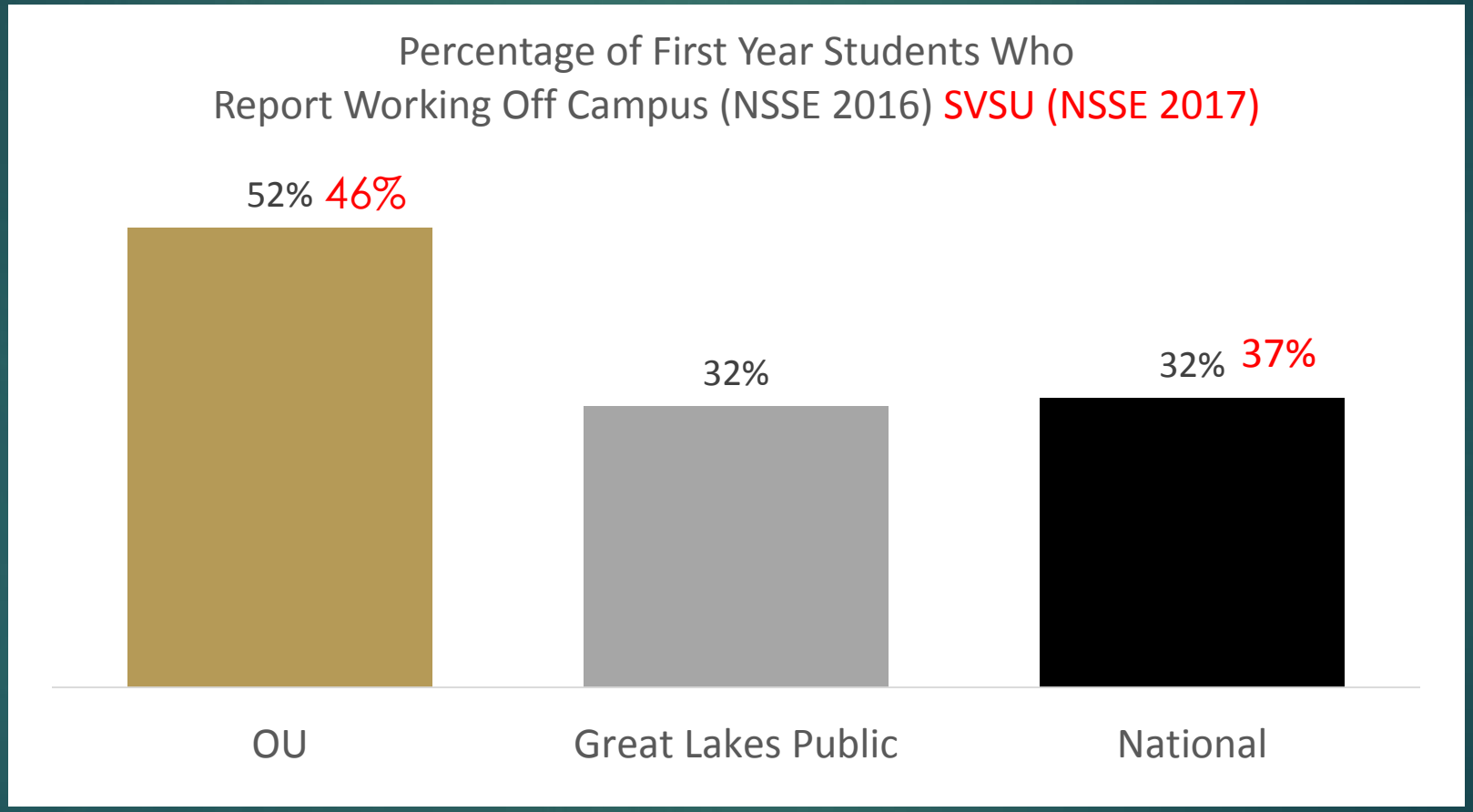
- Generally 3 types from the literature
  - Literal: Exact duplication of the first investigator's sampling procedure, conditions & methods.
  - Operational: When the independent researcher can hopefully follow the same experimental recipe.
  - Constructive: Independent researcher receives nothing more than a clear statement of the outcome from the original author, then purposely avoids imitation of the original work's methods and uses their own to produce an outcome of fact and compare.

So Why are WE talking about this?

- Nick found a study done by Reuben last year interesting!
- He either wanted to say Reuben was brilliant or prove him wrong so he could remind him of that every chance he got at each MI AIR gathering until eternity!
- Can the study be replicated (relatively) well and applied at SVSU?

- Study of Reference: The Role of Unmet Financial Need in Retention (2017 MI AIR).

So Why are WE talking about this?



Refresher: Measurement Issues!

- For many SISs, COA is a variable that can be attached to an individual student's record. **- True**
- But we couldn't use that variable. – **I used the estimation**
- Why?
  - Because, tuition, a huge component of COA, was estimated.
  - A student taking 18 credits was getting the same tuition estimate as a student with 14 credits. But our tuition is based entirely on credits!
- In addition, COA estimates sometime change if the student indicates they are attending only one semester.
  - So the data needed further cleaning to make sure that COA estimates were comparable for students that attended for one semester vs. those that attended for the whole year. **- True**
  - All COAs were recalculated for students based on actual credits. **– See above…all estimated COAs.**
  - Students that attended only for the fall semester had their COA's doubled (as well as their disbursed aid) to keep them in relative sync with students that attended for both fall and winter. **- True**

More Measurement Issues – The Saga Continues

- Expected family contribution is an important part of the formula used here.
- What about students that did not complete the FAFSA?
  - It was then assumed that these students had no unmet need.
  - Is that a true statement? Probably not.
    - But it's hard to assume otherwise.

SVSU's Sample

- Examined 2016 FTIAC Class
- Full time only
- Excluded International Students and Out of State. Not many of them and missing a lot of COA info.
- Final sample was 1,285 (5% of population excluded).
- Have found similar results in other institutional studies: there are complicated relationships between unmet need, student characteristics, and student outcomes.
- It's not going to be as simple as 'if we give these students more money, our retention problems will go away'.
- However, the question remains: if we did address their financial struggles, how much improvement would we expect to see in first year retention rates?

Variables – Completely Replicated and Reproduced

- Outcome of Interest
  - First Year Retention (Binary)
- Control Variables
  - Unmet Need (in thousands of USD)
  - Expected Family Contribution (in thousands of USD)
  - ACT Composite Scores
  - High School GPAs
  - First Term Credits
  - Housing Indicator (binary yes/no)
  - First Generation Status (binary yes/no)
  - Underrepresented Minority Stats (binary yes/no)

Results

Variable - Oakland	B	S.E.	Wald	Sig.	Exp(B)
Unmet Need	-0.091	0.010	78.6520	0.0000	0.9130
EFC	-0.007	0.003	6.0900	0.0140	0.9930
ACT	-0.021	0.018	1.3430	0.2470	0.9800
HS GPA	1.044	0.147	50.7600	0.0000	2.8410
First Term Credits	0.124	0.034	12.9720	0.0000	1.1320
Housing Indicator	-0.116	0.116	0.9980	0.3180	0.8900
First Generation Indicator	-0.001	0.154	0.0000	0.9940	0.9990
URM Status	0.258	0.141	3.3400	0.0680	1.2950
Constant	-3.209	0.677	22.4830	0.0000	0.0400
Nagelkerke R2	0.192				

Variable - SVSU	B	S.E.	Wald	Sig.	Exp(B)
Unmet Need	-0.0001	0.000	41.132	0.000	1.000
EFC	-0.00001	0.000	1.692	0.193	1.000
ACT	-0.00439	0.023	0.001	0.982	0.999
HS GPA	1.04778	0.170	37.685	0.000	2.843
First Term Credits	0.09788	0.061	2.657	0.103	1.104
Housing Indicator	0.19976	0.154	0.291	0.590	1.087
First Generation Indicator	0.0059	0.154	0.008	0.928	0.986
URM Status	0.03857	0.179	0.000	0.983	1.004
Constant	-3.34804	0.928	13.971	0.000	0.031
Nagelkerke R2	0.133				

It might look like the beta weights for OU and SVSU are vastly different for Unmet Need and EFC, but they are actually very similar. OU's data is in thousands of USDs and SVSU's data is in dollars.

Logistic Regression Math

- Retention Rate = $F(x) = \frac{1}{1 + e^{-(B_0 + B_1 x + B_2 x + \cdots)}}$

- Those beta weights correspond to the B column in the results slide. All you have to do is plug them in and go.

- In order to get our final answer though, we have to run some estimates on what the average might be for the other variables.
  - Some variables, like first generation, have basically no impact on the outcome at all.
  - But you can run a number of different scenarios to probe how the retention rate changes for various student demographics.

Outcome?

- Each $1000 that we decrease Unmet Need by, retention rates increase by about 1.74%. In the OU study it was 1.6%
- This finding is very consistent with previous research that suggests that every $1000 equates to somewhere between 1% and 2% change in retention rates.
  - Includes other regression studies
  - Includes Regression Discontinuity studies
  - Includes Interrupted Time Series studies
- Even if we add $5000 of additional aid to the most needy group, we would expect their retention rates to improve from 50% to less than 60%. – Almost identical findings here at SVSU.

Replication & Reproducibility: Final Thoughts

- Replication and Reproducibility lie at the heart of true science: Need evidence of reliable effects
- Findings need validation and should be challenged
- Today's Example shows one way to measure generalizability across institutions. We all work on similar issues, challenges and inquiry
- Reproducing research provides a great way to collaborate and can initiate new learning (from methodology to outcomes)
- May lead to new discovered theories that have yet to be discussed in the literature

Questions?

Contact Information:

Nick Wagner

njwagner@svsu.edu

989-964-2468

Reuben Ternes

ternes@oakland.edu

248-370-2559