the WINNOWER

SOCIAL SCIENCES

[17] No-way Interactions

URI SIMONSOHN

CORRESPONDENCE:

DATE RECEIVED:
June 10, 2015

DOI:
10.15200/winn.142559.90552

ARCHIVED:
March 05, 2015

CITATION:
Uri Simonsohn, [17] No-way Interactions, *The Winnower* 2:e142559.90552 , 2015 , DOI: 10.15200/winn.142559.90552

This post shares a shocking and counterintuitive fact about studies looking at interactions where effects are predicted to get smaller *(attenuated* interactions).

I needed a working example and went with Fritz Strack et al.'s (1988, .pdf) famous paper [933 Google cites], in which participants rated cartoons as funnier if they saw them while holding a pen with their lips (inhibiting smiles) vs. their teeth (facilitating them).

The paper relies on a sensible and common tactic: Show the effect in Study 1. Then in Study 2 show that a moderator makes it go away or get smaller. Their Study 2 tested if the pen effect got smaller when it was held only *after* seeing the cartoons (but before rating them).

In hypothesis-testing terms the tactic is:

Study	Statistical Test	Example
#1	Simple effect	People rate cartoons as funnier with pen held in their teeth vs. lips.
#2	Two-way interaction	But less so if they hold pen *after* seeing cartoons

This post's punch line:
To obtain the same level of power as in Study 1, Study 2 needsat least twice as many subjects, per cell, as Study 1.

Power discussions get muddied by uncertainty about effect size. The **blue fact** is free of this problem: *whatever* power Study 1 had, at least twice as many subjects are needed in Study 2, per cell, to maintain it. We know this because we are testing the reduction of that *same effect*.

Study 1 with the cartoons had n=31 per-cell.1 Study 2 hence needed to increase to at least n=62 per cell, but instead the authors decreased it to n=21. We should not make much of the fact that the interaction was not significant in Study 2

(Strack et al. do, interpreting the *n.s.* result as accepting the null of no-effect and hence as evidence for their theory).

The math behind the **blue fact** is simple enough (see math derivations .pdf | R simulations| Excel

Simulations).
Let's focus on consequences.

A multiplicative bummer
Twice as many subjects per cell sounds bad. But it is worse than it sounds. If Study 1 is a simple two-cell design, Study 2 typically has at least four (2×2 design).
If Study 1 had **100** subjects total (n=50 per cell), Study 2 needs at least 50 x 2 x 4=**400** subjects total.
If Study 2 instead tests a three-way interaction (attenuation of an attenuated effect), it needs N=50 x 2 x2 x 8=**1600** subjects .

With between subject designs, two-way interactions are ambitious. Three-ways are more like no-way.

How bad is it to ignore this?
VERY.
Running Study 2 with the same per-cell *n* as Study 1 lowers power by ~1/3.
If Study 1 had 80% power, Study 2 would have 51%.

Why do you keep saying at least?
Because I have assumed the moderator *eliminates* the effect. If it merely *reduces* it, things get worse. Fast. If the effect drops in 70%, instead of 100%, you need FOUR times as many subjects in Study 2, again, per cell. If two-cell Study 1 has **100** total subjects, 2×2 Study 2 needs **800.**

How come so many interaction studies have worked?
In order of speculated likelihood:

1) *p*-hacking: many interactions are post-dicted *"Bummer, p=.14. Do a median split on father's age… p=.048, nailed it!"* or if predicted, obtained by dropping subjects, measures, or conditions.

2) Bad inferences: Very often people conclude an interaction 'worked' if one effect is $p<.05$ and the other isn't. Bad reasoning allows underpowered studies to "work."
(Gelman & Stern explain the fallacy .pdf, Nieuwenhuis et al document it's common .pdf)

3) Cross-overs: Some studies examine if an effect reverses rather than merely goes away,those may need only 30%-50% more subjects per cell.

4) Stuff happens: even if power is just 20%, 1 in 5 studies will work

5) Bigger ns: Perhaps some interaction studies have run twice as many subjects per cell as Study 1s, or Study 1 was so high-powered that not doubling n still lead to decent power.

1. Study 1 was a three-cell design, with a pen-in-hand control condition in the middle. Statistical power of a linear trend with three n=30 cells is virtually identical to a t-test on the high-vs-low cells with n=30. The **blue fact** applies to the cartoons paper all the same. [↩]

- Twitter
- Facebook

- More
-