IBTN 2020 GLOBAL

ibtn
international
behavioural
trials network

Pilot Trials in Health-Related Behavioral Intervention Research: Problems, Solutions, and Recommendations

Kenneth E. Freedland, PhD
Washington University School of Medicine
St. Louis, Missouri USA

Symposium on Conducting Pilot and Feasibility Trials
Chris Noone, PhD, Chair
May 29, 2020

Disclosure

- Research funding: National Heart, Lung, and Blood Institute (USA)
- Editorial stipend: Society for Health Psychology
- This talk is based on:
  - Freedland KE. Pilot Trials in Health-Related Behavioral Intervention Research: Problems, Solutions, and Recommendations. *Health Psychology*, in press.

“Preliminary Efficacy”

- Across diverse areas of intervention research, pilot studies have traditionally been designed to serve as “preliminary efficacy” trials (PETs).
  - PETs are miniature RCTs that are *intentionally* underpowered & typically *severely* underpowered.
  - I’m not talking about trials that were supposed to be adequately powered but that were plagued by under-enrollment.

"Preliminary Efficacy"

- Many PETs yield disappointing efficacy results, thereby stifling work not only on useless interventions but on promising ones as well.
- PETs with pleasing results are often taken as green lights for larger efficacy trial proposals.
- PET effect sizes are often used in power analyses for proposed RCTs.

Serious Drawbacks

- Only a very small proportion of published PETs have ever led to a full-fledged RCT; probably <10%.
- Positive results do *not* guarantee that a larger RCT will also yield positive results.
  - To the contrary, many small trials with positive results have been followed by larger trials with *negative* results.
- Despite their glaring weaknesses and lack of replication, PETs are often said to have important implications for clinical practice.

Serious Drawbacks

- PETs do much more harm than good for the cause of evidence-based behavioral medicine.
- Yet many researchers and reviewers still believe:
  - Pilot studies are *supposed* to serve as PETs.
  - RCT proposals should (or must) include positive preliminary efficacy findings.
  - PET reports add to rather than detract from the research literature.

Serious Drawbacks

- PETs are also problematic for journal editors.
- Cover letters often sound to us like this:
  - "We're submitting a severely underpowered PET, but please take it seriously as if it were a full-fledged RCT. And please give us a break – it's only a pilot study, after all, so don't hold it to the same standards as a full-fledged RCT."

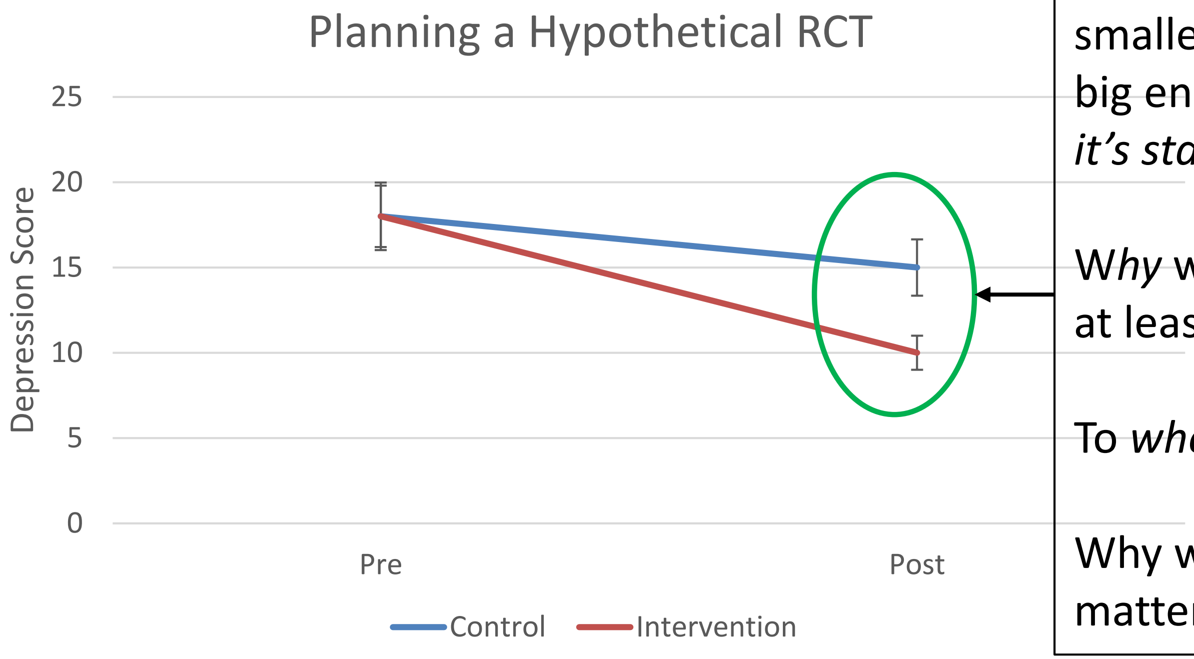
Are There Any Alternatives?

- Helena Kraemer and her colleagues have written eloquently about why we should not use PET effect sizes in RCT power analyses.
- What should we do instead?
- And how can we convince reviewers to fund our RCT proposals if we don’t give them some pleasing PET data?

Alternative Effect Size

- Kraemer et al. argue that instead of using a PET effect size, we should define and justify a *threshold of clinical significance* (TCS).
  - TCS is the smallest between-group effect size that would matter in some meaningful way, e.g. an effect as large as the TCS would be big enough to:
    - disturb clinical equipoise among experts in the field,
    - encourage further research (e.g., a larger trial), or
    - justify changes in clinical practice.

Threshold of Clinical Significance



TCS: if the difference is smaller than this, it won't be big enough to matter, *even if it's statistically significant.*

Why would an effect that's at least this large matter?

To *whom* would it matter?

Why would anything less *not* matter to them?

Threshold of Clinical Significance

- In general, PETs can’t answer those sorts of questions, so they’re irrelevant for defining a TCS.
- Meaningful TCS values may be based on:
  - Epidemiological and/or clinical research findings;
  - Long-term goals of a clinical research program;
  - Stakeholder and cost-effectiveness perspectives;
  - Needs assessments; etc.

IBTN 2020 GLOBAL

Threshold of Clinical Significance

ibtn
international behavioural trials network

- Kraemer et al. recommend that we use a particular effect size index, the *success rate difference* (SRD), to define TCSs for our trials.
  - Use SRD in the power analysis and for between-group comparisons.
  - Translates into more familiar / interpretable indices including number needed to treat (NNT).
- See: Kraemer, H. C., Neri, E., & Spiegel, D. (2020). Wrangling with p-values versus effect sizes to improve medical decision-making: A tutorial. *Int J Eat Disord, 53*(2), 302-308.

Success Rate Difference

- SRD is remarkably versatile; can be used with continuous, categorical, and time-to-event outcomes.
- T and C are two randomly sampled patients from the Treatment and Control arms.
  - The “success” is the one with the better outcome.
  - SRD roughly equivalent to $p(T_s – C_s)$.
  - If every success is a T, then SRD =   1.
  - If every success is a C, then SRD = -1.
  - If there’s no difference, then SRD = 0.
  - Trial planners usually expect or hope for 0 < SRD ≤1.

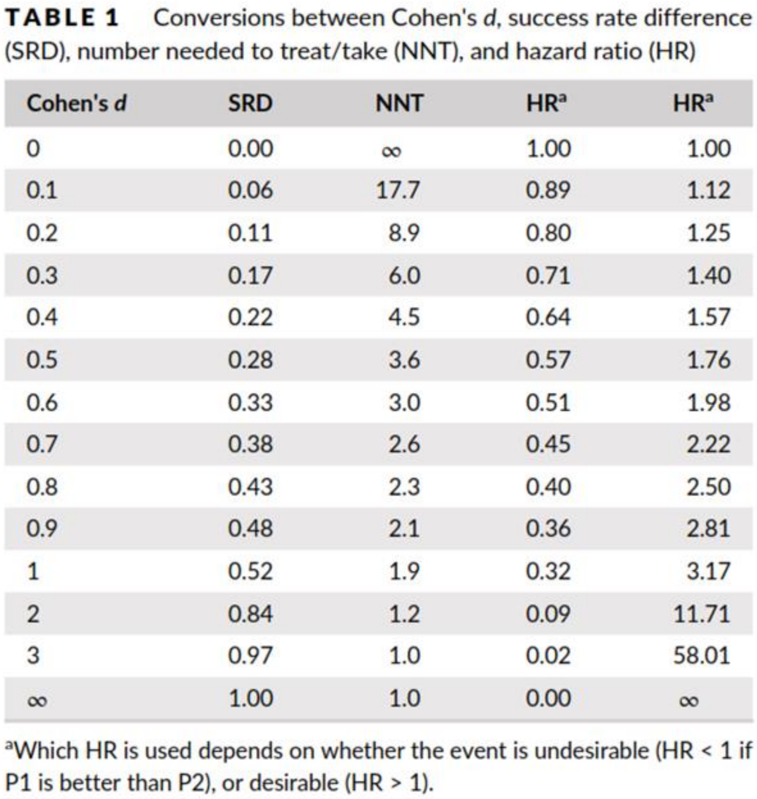


TABLE 1 Conversions between Cohen's *d*, success rate difference (SRD), number needed to treat/take (NNT), and hazard ratio (HR)

Cohen's *d*	SRD	NNT	HR[a]	HR[a]
0	0.00	∞	1.00	1.00
0.1	0.06	17.7	0.89	1.12
0.2	0.11	8.9	0.80	1.25
0.3	0.17	6.0	0.71	1.40
0.4	0.22	4.5	0.64	1.57
0.5	0.28	3.6	0.57	1.76
0.6	0.33	3.0	0.51	1.98
0.7	0.38	2.6	0.45	2.22
0.8	0.43	2.3	0.40	2.50
0.9	0.48	2.1	0.36	2.81
1	0.52	1.9	0.32	3.17
2	0.84	1.2	0.09	11.71
3	0.97	1.0	0.02	58.01
∞	1.00	1.0	0.00	∞

[a] Which HR is used depends on whether the event is undesirable (HR < 1 if P1 is better than P2), or desirable (HR > 1).

Plausibility

- By defining a TCS in your RCT grant proposal, you're telling the reviewers:
  - The efficacy effect size will have to be *at least* that large for the results to have much of an impact.
    - The desired impact(s) depend on the goals and phase of the research program and the primary purpose or specific aims of the study.
  - You won't conclude that the intervention is efficacious unless the observed effect is as large as or larger than the TCS; statistical significance is not enough.
- Consequently, the reviewers will want to know why you think there's a good chance that the observed effect will turn out to be at least that large.
  - How do you do that without giving them any "preliminary efficacy" data???
- You'll have to persuade them that an effect as large as the TCS is a *plausible* or *very plausible* outcome of the proposed trial.

Plausibility

- PETs aren’t designed to make plausibility arguments.
- The implicit (faulty) logic is usually something like this:
    - Our PET showed that our treatment is efficacious.
    - It must be *very* efficacious for us to have found such a big effect in such a small trial.
    - We need to do a big trial just to confirm that our treatment is as good as we’ve already demonstrated.
    - The trial will almost certainly show positive results because our PET showed positive results.

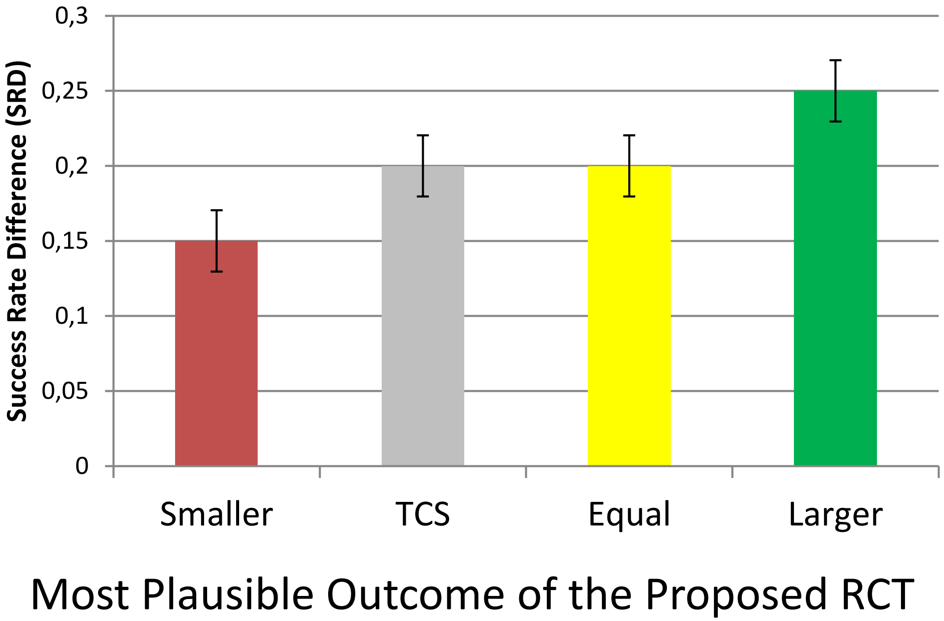
Plausibility

- Something is plausible if it seems
  - truthful, reasonable, credible, believable (dictionary.com)
  - likely to be true; believable (Cambridge Dictionary)
- Asserting that an expected effect is a *plausible* outcome of a proposed trial is very different than arguing that it's essentially a sure thing.
  - "It's basically a sure thing" is the implicit argument when PETs are used to justify proposed RCTs.
- The principle of equipoise suggests that it would be unethical and pointless to conduct a clinical trial if there were little or no doubt about how it would turn out.
- Like recruiting a bunch of skydivers, randomizing them to parachute vs. no parachute groups, and comparing their mortality rates to see if parachutes work.
- Preliminary data can't *guarantee* that an RCT will turn out well; at best, it can be used to argue that a favorable outcome is *plausible* or *very plausible.*

Plausibility

- Evidence of plausibility can come from:
  - Scientific credibility of the intervention, e.g.,
    - Intervention is grounded in basic science and research is embedded in a translational or optimization framework (e.g., MOST, ORBIT, SOBC, etc.)
    - Intervention has undergone systematic development, testing, refinement.
  - Literature showing that the TCS is incrementally – not dramatically -- better than what's been achieved in previous studies, by similar interventions, etc.
  - Proof-of-concept data from uncontrolled trials suggesting that an effect as large as the TCS is probably attainable.
- Beyond the supporting evidence, an effect as large as the TCS won't seem plausible unless the reviewers think that the trial is well designed and feasible, and that it will be rigorously conducted.

Plausibility



Reviewers might not like your proposal unless the TCS makes sense to them AND an effect as large or larger than the TCS seems like a plausible outcome of the trial.

Making the Case for an RCT

- Define the TCS for your proposed trial.
- Convince the reviewers that it’s well chosen, i.e., it’s an effect that would be worth finding.
- Convince the reviewers that you have a reasonably good chance of finding it.

Making the Case for an RCT

- Proper pilot studies (ones that evaluate the feasibility of an RCT) can be very valuable, but:
  - They won’t help you define the TCS for your RCT.
  - They won’t provide much of the evidence you’ll need to support your plausibility argument.
  - They *shouldn’t* provide preliminary efficacy data.

Whither Preliminary Studies?

- Translational research and optimization frameworks (e.g., ORBIT, MOST, SOBC) show that a variety of different kinds of studies can help to lay the groundwork for RCTs.
  - In general, these frameworks don’t recommend PETs.
  - Contrary to the traditional (and still popular) belief, PETs are not the be-all and end-all of preliminary research on behavioral interventions.
  - We’ll be much better off without them.

The Battle is Joined, And It's Not Over Yet!

- Some grant reviewers know PETs are problematic.
- Others don't; some may want, encourage, or expect you to provide PET data.
  - You've got to know when to hold 'em,
    Know when to fold 'em,
    Know when to walk away,
    And know when to run…
    - Kenny Rogers, *The Gambler*
- Please help educate reviewers, colleagues, and trainees that PETs are obsolete & counterproductive *and* that there are better alternatives.