Reviewer's report

Title: Analysis and reporting of stepped-wedge randomised-controlled trials: synthesis and critical appraisal of published studies, 2010-2014

Version:1**Date:**19 March 2015

Reviewer:J S Gruber

Reviewer's report:

- Major Compulsory Revisions

NONE

- Minor Essential Revisions

1. The authors should review the PRISMA statement for the reporting of systematic reviews (and meta-analyses) and incorporate these reporting recommendations into their manuscript.

2. PAGE 14, "REPORTING", 2nd Paragraph: " Reporting both balance between groups and balance between conditions might be advisable so as to identify imbalances arising by chance as well as by secular changes."

The authors should be more explicit about what they are referring to here by "group". Is this cross-over group (when multiple groups crossover at the same time point), or cluster; we would not expect clusters to be balanced -- I also do not believe we would expect cross-over groups to be balanced. Since time is effectively randomized in SWT (or clusters are randomized over time), we would expect successful randomization to balance baseline cluster/individual covariates over time -- that is, on average treatment and control time periods will be balanced with respect to baseline characteristics. Given this, "between conditions" seems the most appropriate method to report balance metrics – given that the fact that clusters spend different amounts of time in each condition is accounted for. The authors should elaborate and/or justify this statement here or elsewhere in the manuscript.

3. PAGE 14, "ANALYSIS", 2nd Paragraph: This is a good discussion of the potential for the intervention to change effectiveness over time. However, I read the authors' discussion of the issue here, and earlier in the paper, as a the change in effectiveness as solely a potential source of bias (which it is). However, one of the advantages of the SWT over a parallel arm CRT is the potential to actually evaluate "time on intervention" as a primary parameter of interest (not possible in a parallel arm trial). A discussion of this potential would be a nice addition to the paper, and a great contribution to the literature.

4. PAGE 15, "ANALYSIS", 3rd Paragraph "This is not to be recommended: inclusion of data before and after the crossover in this way incorporates a

before-after comparison that is uncontrolled, and may cause bias in the intervention effect estimate."

The authors should further justify this. Is this true in all cases? Doesn't including data from these periods increase precision of estimates by borrowing information from these time periods in a regression model (when all units are in Tx or Control)? The analog would be a CRT with a closed cohort and repeated measures and a complete baseline (untreated) for both tx and control arms. It is not uncommon to borrow outcome information from these baseline measures as additional control observations. The authors should further justify this statement.

A potential suggestion for investigators is to conduct sensitivity analyses -- where analyses are presented with and without the first and last steps included if the investigator does not want to drop these data.

5. PAGE 15, "ANALYSIS", 4th Paragraph: "Some studies included relatively few clusters but standard analyses based on mixed effect models or GEE were used, even though these are known to be problematic."

The authors should elaborate on this point, and justify/offer support; it would also be helpful to be explicit about what is meant by "problematic" (which is vague) and offer suggestions for minimum number of experimental units at which these methods are no longer problematic.

6. PAGE 15, "ANALYSIS", Last Paragraph: " To generate evidence against the null hypothesis, researchers can permute the allocation of clusters to groups according to the rules of the randomisation (e.g. stratification), calculating an 'effect' under each permutation, and locate the empirical effect estimate in the distribution of effects estimated under random chance"

This is a great addition to the paper. The authors should consider providing a citation from literature on permutations tests and/or published examples (from CRT) where this was done as a guide .

7. TABLE 2: In the text, or as a footnote to this table the authors should provide more detail on how they assessed risk of lag and risk of fidelity. It was not clear from this table, or the methods.

- Discretionary Revisions

8. PAGE 14, "REPORTING", 3rd Paragraph: a table showing the key "CRT CONSORT" reporting items that are either new for SWT, or need to be adjusted/adapted for the SWT -- perhaps with a brief description of options or avenues for future research would be helpful and an asset to the paper as a future reference. (FOLLOW UP NOTE: Panel 1 found – however recommended table would provide more/clearer information).

Level of interest: An article of outstanding merit and interest in its field

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:

I am the first author on one of the papers reviewed in this manuscript.
I have no other competing interests.