Reviewer’s report

Title: A Reliability assessment of a direct-observation park evaluation tool: the Parks, activity and recreation among kids (PARKS) Tool

Version: 1 Date: 12 February 2015

Reviewer: Sonja A. Wilhelm Stanis

Reviewer’s report:

While the overall idea of the paper and topic is important, there are numerous concerns that need to be addressed. Thank you for the opportunity to review this manuscript.

MAJOR COMPULSORY REVISIONS

1) I found the development of a new tool for this study to be counter to the argument being made for assessing the reliability and validity of tools in a variety of contexts. Although some of the previous tools had limitations in their testing as mentioned (such as small sample sizes, or not tested with youth), given the argument laid out it seems as though taking one of those tools and expanding the testing to other contexts (additional parks, different geographic locations, different age groups) would better meet that objective than developing another new tool. Indeed, the tool development in the methods sections specifically states this new tool was created to be adapted to the Canadian context. Further, the goal of the Kaczynski et al tool [20] is incorrectly stated, as it was beyond just increasing awareness of park features among community stakeholders (and it is unclear why this as one of the goals would be a limitation).

2) Even if one study has not compared reliability across different contexts for a particular tool, several tools have been used in different geographic locations. I am not sure if reliability assessments have been examined or presented in each publication, but that would speak to the reliability across contexts as well. I think it would be good to at least investigate this in the literature a little if not already done.

3) Third paragraph of the background and table 1. I was surprised to see SOPARC listed as a park evaluation (audit) tool. While it is a direct observation tool, it is used to observe physical activity. Certain characteristics (such as if the area is supervised, accessible, etc.) are included to help interpret the activity observations, but the focus of the tool is not to fully assess park characteristics.

4) Methods: I would like to see significantly more detail/clarity regarding the tool development and description. It would be helpful to know the specific number of items from both the POST and BRAT-DO (it may help to integrate the 2nd and 3rd paragraphs in this section). In addition, several places it is stated that new items were adapted to assess features not on either the POST or BRAT-DO. Adapted from what instrument? Or, should “adapted” be changed to “included”? What type of response options were used for each item (e.g., purely
present/absent, or other scales)? What kind of revisions were made based on the pilot testing? Were any of these changes made to the POST or BRAT-DO items? Although not mentioned in this section, it is stated in the Observer Training that the tool was also revised and adjusted after each on-site training day. This appears to be part of the tool development as well (and again, were these to any POST or BRAT-DO items)?

5) If items were pulled from both the POST and BRAT-DO, why was a similar comparison not made to previous BRAT-DO reliabilities in other contexts as well? From table 2, it looks like 17 items were used in both the PARK and POST. So, only 17 of the 92 total items were compared across geographic contexts? I think this needs to be made very clear that it is a small subset of the overall questions that were assessed (including in the discussion).

6) Methods: please change the subheading “Sampling Plan” to “Park Identification”, or perhaps “Park Identification and Sampling” as that better reflects what is presented in that section. In addition, move the sampling plan information in the Results, to this methods section as that information regarding the number of parks assessed and sampling pertains to methods. In this section, also describe what was defined as a “park”. Did it have to have facilities? Or, could it be simply a natural area/open space even if there are not paths, facilities etc.? Would something such as a school playground be included? It would be helpful to know what type of park/open spaces are included in the audits (and that could have implications for what can be assessed through CanMap as well).

7) Observer training: provide more information regarding the observers and training. How were they recruited and who are they? (Ages? Youth? University students? Community members?). Were each of the nine observation training days full days of training? And, to be able to say this is a reliable tool in the conclusion, a qualification regarding the extent of the training needed to correspond with its use is necessary.

8) Reliability assessment of observers: expand this section as it is currently confusing and missing necessary information. (1) Regarding the inter-rater reliability, describe this process more. Were the pairs conducted on the same day? Did they go through the park together or independently? Did every park audited have two observers? Were the pairs always the same two individuals, or were they mixed up? (2) Likewise, provide more details regarding the intra-rater reliability: I found the sentence in this section confusing, and would like to see it re-worded and expanded. It would help to clarify here that this was done through test-retest methods. Also, provide the details included in later sections regarding the time span between the assessments as this is relevant methods information. Further, in the study limitations it is stated that it was not possible to assess intra-rater reliability for all items – it is not clear why this is the case and should be explained in the methods. (3) It is unclear which tests took place during the one week in June 2008 and thus more details and clarification is needed regarding dates. Is this for the inter-rater reliability (in that all 576 parks were audited by two observers in one week in June 2008, and then the intra-rater reliability was assessed through Dec 2008, and in 2009 and 2010?). Or, is this referring to a third assessment against the gold standard stated in the next
sentence? This final assessment also needs further explanation: was this with a subset of the audits (just those that occurred that one week in June) and how many? Further, why is this one sentence the only time this gold standard assessment is discussed (not in the results, etc.), other than a discussion of comparing to a gold standard in the training (is this what that is referring to? Were park audits conducted as part of the training also included as part of the sample?)

9) There needs to be better correspondence between the objectives, methods and results presented. Objectives 2 and 3 are not discussed in the analysis – how were those comparisons assessed? Regarding objective 2, in the results it is stated that they were compared and found to be of similar magnitude. How were they compared? Visually? To a set standard? More is needed on how this was done and which items. The methods also needs more information the POST data (number of parks, where, when). Also, more is needed in the results regarding this (in addition to referring to table 2, describe some of the findings in the text). Regarding objective 3, how was this assessed? There is also no results presented for objective 3, which indicates it really should not be a main objective of this paper.

10) I am confused by the use of additional files for some of the main results (additional file 2 and additional file 3 are the results for objective 1). It seems they should be presented as a table in the main paper, perhaps in a condensed form such as showing ranges for each section, and then refer the reader to the additional file for the detailed table showing all items. Otherwise, why is the data in table 2 more “important” than the other results?

11) A few of the suggestions in the discussion could use more detail. For example, it is stated that some items may need to be modified - in what ways? If including a more objective benchmark, such as what? What would enhanced observer training include (beyond the already 9 day training)?

12) As mentioned above, I am a bit concerned regarding the discussion around objectives 2 and 3 given the few details provided in the methods and results. Further, in the discussion regarding the comparison of the PARK and POST, it is mentioned that there were a few items that had considerably different reliability estimates. These few items were not identified in the results section. And, once again, what is the standard used to determine what is similar and what is different. For objective 3, given that there was no discussion in the methods or results, the discussion of the comparison to desktop park evaluations does not seem appropriate, at least as a major objective. However, there is some results presented in the discussion that could be used to provide results (number of parks not pre-identified, number of parks that were modified, etc.). Further, this article might be relevant to this discussion:

13) Conclusion: Although the lack of validity testing is mentioned in the study limitations, the conclusion that the tool is recommended for use is concerning without some indication of validity. I also think that the conclusion needs to stress the importance of training, as it is through a fairly intensive training that the reliability was established. Future research should also expand beyond just the
reliability of the same items shared between the POST and PARK, but also other items in the PARK as well.

MINOR ESSENTIAL REVISIONS

1) The purpose of the study is worded slightly differently in the abstract compared to the main text, and I interpret them as slightly different meanings. Please edit so that they are the same.

2) First paragraph in background: include a citation for the remainder of the 2nd sentence regarding the adult morbidities.

3) Second paragraph of the background. Be more specific regarding what is meant by park improvements. Specifically, in the last sentence, the study given as an example [11] examined adding fitness zone equipment, which is very different than other possible park improvements (e.g., maintenance repairs, etc.).

4) Table 1: Please include the citation numbers for each reference in table 1 so that readers can more easily link statements in the manuscript to the corresponding tool in the table.

5) I would also like to see a different citation for the POST study, as currently it is only a link to the tool. Given the importance of the tool for this manuscript, please cite a study that includes more background on the tool, and the source of the information in table 1.

6) Sixth paragraph of the background: regarding the statement “there is a need for improved reliability of direct observation audits of park characteristics” seems to be implying that the reliability results for the previous tools are unsatisfactory – is that the intention?

7) Observer training: Move the second sentence regarding the audits being embedded in a larger study up to the previous section (Park Identification and Sampling), as it does not flow well in this section.

DISCRETIONARY REVISIONS

1) First paragraph in the background, “Canadian youth lag far behind other nations”: How far behind? Maybe better to just say “Canadian youth lag behind other nations…”.

2) I recommend cutting back the description of the broader study context some, particularly regarding the other measures (interviews, questionnaires, etc.). It actually creates more confusion regarding what is relevant to this element of the study and what is not.

**Level of interest:** An article of importance 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 declare that I have no competing interests