Author's response to reviews

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

Authors:

Madeleine E Bird (madeleine.bird@umontreal.ca)
Geetanjali D Datta (geetanjali.datta@umontreal.ca)
Andraea van Hulst (a.van.hulst@umontreal.ca)
Yan Kestens (yan.kestens@umontreal.ca)
Tracie A Barnett (tracie.barnett@laf.inrs.ca)

Version: 4
Date: 11 August 2015

Author's response to reviews: see over
Ms. Natalie Pafitis, Executive Editor
BMC Public Health

Tuesday 11 August 2015

Dear Ms. Pafitis,

Please find enclosed our third revised manuscript MS: 2659124091420558, A Reliability assessment of a direct-observation park evaluation tool: the Parks, activity and recreation among kids (PARKS) Tool. We revised the manuscript according to reviewers’ comments and provided a detailed, point-by-point reply to address all comments and outlined the changes. The point-by-point reply to the reviewer is enclosed with this letter.

We look forward to your editorial response.

Sincerely yours,

Madeleine Bird, M.Sc.
Ph.D. Candidate, Epidemiology
School of Public Health, University of Montreal
Research Centre, CHU Sainte-Justine
5757 Decelles avenue, Suite 100
Montreal, Quebec H3S 2C3
Phone: 514-345-4931 extension 5211
Email: madeleine.bird@umontreal.ca
Reviewer's report

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

Version: 3 Date: 22 July 2015

Reviewer: Sonja A. Wilhelm Stanis

Reviewer's report:
Once again, the paper has been greatly improved with additional details and clarity. However, reading through the paper again, I still have a few remaining questions/concerns. With these final changes and clarifications, I believe the paper will be much stronger for publication.

Major Compulsory Revisions

1. Pg. 9, paragraph starting on line 202. I am still confused if these items that were changed are considered as part of the 24 items that were pulled from the POST and BRAT-DO, or part of the 13 new items as several items were consolidated into one.

These changes are made to the items that were pulled from the POST and BRAT-DO. The first sentence of the paragraph formally stated: “During pilot testing, a number of changes were made to the POST and BRAT-DO items that were adapted to the PARK tool.” In order to improve clarity, the sentence has been changed to the following: “During pilot testing, items drawn from the POST and BRAT-DO were changed and adapted to the PARK tool.” (Line 202-203).

2. P. 15, line 285-286. Although the issue regarding the time between intra-rater reliability checks is mentioned in the limitations, I do not feel that is sufficient for this concern. Specifically, this is not a valid intra-rater reliability check given that time lapse. Instead, with that amount of time, it is instead a check if the environment has changed or not. Indeed, on page 16, over several months to a year, it is reasonable that graffiti, litter, etc. will indeed change. I recommend sending a much lower threshold for which pairs can be included (perhaps 2 months which I would still argue is long) in this analysis. While that will reduce the number of checks that can be included, it will make more conceptual sense to be treated as an intra-rater reliability check. (Also note that the methods on page 13 reports the median number of days, while the discussion on page 21 reports the mean number of days).

Thank you for this comment. We agree that the time delay could increase the likelihood that the environment has changed. However, after careful consideration of our particular set of circumstances and analyses we believe that the consequence of this would be that our results would be an underestimation of
the test-retest reliability estimates due to non-differential misclassification. Therefore, we remain confident that the reliability assessment is indeed valid. We explain our reasoning below.

The test-retest is interested in how consistently examinees respond to a test at different times. A critical question in the design of a test-retest study is the length of time that should elapse between tests. It is clear the reviewer is well aware that there is no hard and fast rule about when in time a retest should be undertaken. In order to think through the reviewer’s comment, we went back to a standard text Crocker & Algina’s *Introduction to Classical and Modern Test Theory* (Cengage Learning, Mason, Ohio, 2008) where we found the following regarding the time interval between tests for a test-retest study, “There is no single answer. The time period should be long enough to allow effects of memory or practice to fade, but not so long as to allow maturational or historical changes to occur in the examinees’ true score.” (Chapter 7, page 133.) Depending on the purpose of the test, and the age group participating in the testing, the time between tests can range from as little as one week to several years. Often among adults (as opposed to infants, for example), waiting a longer time between test intervals is desirable to allow effects of memory to fade (Crocker & Algina, 2008, Chapter 7).” In the present test-retest reliability assessment, there was a wide range of times between tests, however this, we argue, does not render the test invalid. To the contrary, we believe the range in time intervals between tests in the present study allowed for sufficient time between tests for the observers to evaluate the parks with “fresh eyes”.

It is certainly true that a small number of items on the tool, such as graffiti, litter and vandalism, will likely change substantively from one test time to another. However, for these types of items, which are inherently transient, the risk for substantive changes are equally true if the time interval were one day, two months, or one year. Graffiti can appear overnight, and litter can be picked up in an afternoon. It is an acknowledged weakness of the test-retest method for these types of items, but it does not invalidate the test as a whole.

The effect the time lapse may have had on the test results, as stated in the limitations, is an underestimation of the test-retest reliability estimates due to non-differential misclassification (which would be greater for the more transient items on the tool). The time lapse did not have an effect on the built environment for the majority of the items on the tool. More over the longer time lapse was likely important to enable a valid test-retest assessment because the evaluators were less likely to remember the first park evaluation. The overall high kappa coefficients and percent agreement results for the test-retest study attest to this statement, and attest to the validity of the methods used. Had we observed low coefficients and agreement between the test and retest occurrences, we could then postulate that one of the reasons for the lack of agreement may have been due to actual changes in the physical environment (as opposed to changes in how raters evaluated the same environment). As such, the agreement was high
and we have no reason to think that this level of agreement was due to environmental changes that may have occurred in the time between evaluations.

Indeed, given this, we have added a sentence in the limitations section regarding the time interval between test-retests (lines 472-474): “Nevertheless, the overall high kappa and percent agreement between test time one and two attest to the general validity of the time interval between the test-retest conducted here.”

We would also like to thank the reviewer for pointing out that our reporting of the mean and median in different places was confusing. We had intended to provide increased clarity by offering two different metrics instead of less clarity. Given that it may be confusing that there are two different types of numbers reported for the time-interval, we have added both statistics in both places for clarity.

3. Pg 16, lines 350-352: I found this validity assessment to have very limited information, particularly compared to the other reliability items. There was no table, no kappa scores, not presented by item. Given the description in the methods on page 13, it seemed this was a significant enough portion of the study to warrant more detail. Alternatively, perhaps the validity check could be removed if that is not the focus.

The reviewer is correct that the focus of the manuscript is not the validity check. Given this comment, we have now removed this minor validity assessment from the manuscript; if the editor prefers to retain it, we can re-insert into the manuscript.

Minor Essential Revisions

1. I disagree with the statement on pg 5 lines 98-99. The CPAT was also developed based on the Bedimo-Rung conceptual model [6] and the BRAT-DO [19].

We have re-read the article describing the CPAT [20] in order to address this comment. This article states that seven existing tools, one of them being the BRAT-DO, were reviewed prior to development of the CPAT, and that the CPAT, “… compares favorably with conceptual models about elements of parks that are important for physical activity,” (p. 246) followed by references to the Bedimo-Rung conceptual model as well as an article from Loukaitou-Sideris & Sideris (2010). It is not stated in the article that the CPAT was developed based on the Bedimo-Rung conceptual model and the BRAT-DO. Does the reviewer know of an additional reference which we might have missed that states that the CPAT is based on the Bedimo-Rung conceptual model? If so, we would be happy to change the text.

2. Pg. 5 line 105: It seems that the reference [26] should instead by [24].

Thank you for pointing this out. The reference has been changed.
3. Pg. 6, line 125 – cite the tool (perhaps 26 here)

The tool has been cited here.

4. Pg. 9 line 205, sentence starting with “For example, in the PARK tool…”. How as this changed after the pilot test?

We apologize, but we are not exactly certain what the reviewer is asking here. Perhaps she is asking what changes were made to the PARK tool following the pilot testing? If so, we hope the text makes it clear that the changes were made during pilot testing with the sentence, 'During pilot testing, items drawn from the POST and BRAT-DO were changed and adapted to the PARK tool.” (Lines 202-203) In hopes of addressing the question, the sentence has been changed to the following (changes highlighted here): "For example, in the PARK tool, unlike in the BRAT-DO and POST, every activity installation item had three adjoining qualifying items added to them: check if the installation is accessible, in good condition, and restricted.” (Lines 204-207)

5. Pg 14, line 305. If 36 items were drawn directly from the POST, why were only 17 comparable? Or, were they not all drawn directly from the POST?

On page 8, line 181, it is stated that, “Almost all of the items [drawn from the POST and BRAT-DO] were moderately to significantly re-worded for the PARK tool and response scales were changed.” Therefore, we were only able to compare those items that were directly comparable, in other words, those items that had not been significantly re-worded, and that were available on the reliability assessment provided on the POST website. In addition, a few items could have been compared because they were worded similarly enough for comparison (such as ‘Interesting for walking’), however reliability assessments were different for these items. The POST used the intraclass correlation coefficient (ICC) for some items to assess reliability, whereas we used kappa coefficients for all reliability assessments, and therefore we could not compare these results. This explanation has now been added to the manuscript (Lines 296-302):

“Thirty-six items on the PARK Tool were drawn directly from the POST; however only seventeen items could be directly compared with items assessed in Montreal, Canada, because a number of items drawn from the POST had been substantially modified for the PARK Tool, and because not all items on the POST used the kappa coefficient to estimate reliability (intraclass correlation coefficient was used instead).” (Lines 297-301)

6. Pg 15, line 336: Is the “presence of pedestrian safety” a set of items? I found this confusing given that all the others were for specific items.
The presence of pedestrian safety is one item, not a set of items. The complete tool is provided in an additional file.

7. Reorder tables. Table 2 (pg 17) is mentioned after tables 3-5 in the text.

The tables have been reordered.

8. Do not include the reference to the satellite images in the conclusion given that is such a limited part of this study, particularly now. I recommend instead moving lines 506-511 to the discussion of this, or simply deleting it if it is repetitive to move it.

The reference has been removed.

Discretionary Revisions

9. Abstract pg 3, line 52: I would remind the reader that this was for only 17 items (Inter-rate reliability estimates of 17 items tested in Montreal and Perth were of similar magnitude)

It has been added here per the reviewer's request.

10. I recommend switching the order of the two subsections on page 14. Specifically, the statistical analysis subsection is just referring to the previous section on page 13 (Observer reliability and validity assessment). The subsection regarding comparison of the POST and PARK breaks up those two related subsections, and includes a somewhat confusing reference to the next section. By moving the statistical analysis section up, the related sections flow better and there is no longer the need to reference information provided in the following section.

We have removed the reference that caused confusion to the reviewer and believe the flow and order of the subsections is now improved.

11. Pg 17, lines 373-376. This was initially confusing to read, and I recommend splitting the two sentences for clarity. “Percent agreement was not available for one item (i.e. xxx) on the POST. For two items (presence of graffiti and litter), the PARK tool had moderate agreement (69.34% and 67.30% agreement, respectively) while the POST had good to excellent agreement (78.26% and 76.00% agreement, respectively) for these same items.”

This sentence has been split in two (lines 366-367).

12. Lines 402-436: this is a very long paragraph that I encourage you to break into at least 2.
13. I recommend combining a few paragraphs in the limitations, as the last two are currently a bit choppy. The paragraph starting on line 486 could be moved to the end of the first paragraph in the limitations, and start with “Further, comparisons of items across….”. Then, the paragraph starting on line 491 could be expanded by moving the last sentence from the conclusion to this paragraph about future research.

These changes have been made to the limitations section.

**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