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: 3
Date: 2 July 2015

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

Thursday 2 July 2015

Dear Ms. Pafitis,

Please find enclosed our second 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 outline the changes. Following the reviewer’s requests, we made significant changes to the methods, and results sections, and added more details where requested. 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
The paper has been greatly improved with additional details and clarity. I still have a few remaining questions/concerns.

1) The need for a new tool is still not adequately described in the manuscript. It still appears that the primary justification is that the current tools haven’t been tested in a variety of contexts. In the response letter the authors commented that the reason for the new tool was that the current tools could not meet the needs of the study – could this be briefly explained in the manuscript? In which ways were those previous tools lacking that this current tool addresses?

The need for the new tool has now been described in the manuscript, lines 124-139: “For the present study, the original intent was to use an existing park direct-observation tool (the Public Open Space Tool, or POST), however it became clear that no single existing tool available at the time of the study (2007) was able to meet the needs of the overall study objectives in terms of efficiency and relevance. At the time of the present study, the two park audit tools [20, 21] that were developed for assessing parks for physical activity among youth were not yet available. The park audits in the present study were embedded in larger detailed neighborhood direct-observation audits of a 500 m walking network buffer around the homes of study participants, therefore requiring thorough yet efficient park evaluations given that the evaluation of the 500 m walking network buffer and the three closest parks near the homes of the youth involved in the study had to be conducted during a one-day visit. The relevance of items was also important regarding the efficiency of the audit, as a number of items on existing tools at the time were not relevant for the study context. Thus, a new study tool, drawing on items and methods from different existing tools, and incorporating new items and response schemes, was developed for the present study. Given these many changes, an independent reliability study was warranted; furthermore, the tool described herein may be useful for replication and application in other studies.”

I think it would also help to directly acknowledge the impact of the timing of the PARK development, as the two audit tools that you acknowledge do focus on park characteristics related to physical activity in youth were published after the date indicated as the start of this current study (2007). Otherwise, I think more is needed to explain why those then were also not sufficient.

The timing of the PARK development tool as it precedes the availability of the two other youth-oriented park audit tools is now directly acknowledged, lines 127-129: “At the time of the present study, the two park audit tools [20, 21] that were developed for assessing parks for physical activity among youth were not yet available.”

I also recommend recognizing those relevant two tools that were not included in the development process of the PARK in the discussion or limitations, and perhaps drawing from elements of those newer tools in any revisions of this tool.

The following recommendation has been added to the limitations section of the manuscript, lines 480-484: “In addition, at the time of this study’s tool development (2007), the other two park audit tools developed for youth activity [20, 21] were not yet available. Any future direct audits of parks for physical activity among youth could consider the items on these other two tools when either developing a new tool or choosing to use an existing one.

2) I am still struggling with the third objective. It doesn’t seem to fit with the rest of the paper as a primary objective. In particular, the identification of a park presence is not actually part of the new audit tool (which focuses on assessing the park characteristics and features). And, it is not
even mentioned in the opening paragraph of the discussion (I would argue because it does not fit). I think it is good to include as a secondary finding that emerged in the process, but I don’t think it is adequately set up or examined to be a primary study objective. Indeed, some of the set up and discussion is confusing around this given that it actually focuses on evaluating park characteristics rather than the identification of park presence (for example pg 18). Instead, I would simply discuss that process and identification discrepancies as part of the methods (rather than a separate research question, brief mention in the background, and short result), and then you can still include a brief discussion regarding the method (as have on pg 19). However, with that, you still may want to include additional details in the methods about the differences in boundaries (and were there any that were pre-identified but then not found?), and if they were a specific type of park that were frequently not identified through CanMap that might give insight into why they are missed?

As per the reviewer’s suggestion, we have removed the third objective as a primary objective and have included it as a secondary finding. This objective has been removed from the abstract, there is a brief mention of it in the background, lines 144-146, and methods, lines 246-247, as well as a brief discussion paragraph.

The discussion paragraph included is the following, lines 450-467:

“A secondary finding regarding the large number of parks that were not identified by GIS warrants a brief discussion. The exercise of pre-identifying parks using geocoded satellite/administrative data prior to observers entering the field facilitated the on-site evaluation process. However, only 64%, of the sample was pre-identified, meaning that observers, using on-site methods, newly identified over a third of the parks evaluated. In addition, some of the parks that were pre-identified were found to have different boundaries when identified on-site. When this was the case, observers would modify the park on the observer map to reflect its actual size or shape. There were no pre-identified parks that were then not found, however a very small number of areas that were pre-identified using GIS were later identified on-site as golf courses or cemeteries. There were no systematic differences between the types of parks identified via GIS or on-site, other than size. There were four very large parks in the sample (> 200 000 m²) that were all identified by GIS. The on-site identification of parks allowed for a more valid sample of the parks of interest in terms of number, location and size, suggesting that studies using satellite images for park identification should validate their findings using on-site verification with a representative sample of the parks. Seeing as direct observation requires significantly more resources than desktop audits, further research should be conducted to work toward improving park identification through satellite images and thus improve capabilities to conduct reliable desktop assessments.”

3) I found the Tool Development section of the methods to still be confusing.

a. I had a hard time following the new section outlining which items came from which tool. I think this is important and needs to be a bit clearer. For instance, it is stated (per my understanding) that 22 items came from both tools, then 23 from the POST and 21 from the BRAT-DO (are these different than the 22? Overlap?). Plus the 8 new items – how does this correspond with the total 92 items? (Might also help to state the 92 total items earlier in this section – maybe even put that last paragraph first, and then describe how that was developed).

We agree that this new section was confusing and we have re-written the paragraph in hopes that it will be clearer. It is now worded as the following, lines 178-191: “The present tool drew the
majority of its items from the POST and the BRAT-DO including 24 items that appear in both tools, 12 items that appear exclusively in the POST and 43 items that appear exclusively in the BRAT-DO. In addition, thirteen items were newly developed, for a total of 92 items. Almost all of the items were moderately to significantly re-worded for the PARK tool and response scales were changed. For example, the POST has an item, “Are picnic tables present?” with response scale 1=yes; 2=no, whereas in the PARK tool, the item is worded as, “Picnic tables”, with response scale 1=yes, in usable condition; 2=yes, but unusable; 3=no. More details on the changes made to items are described below. Most of the qualitative items about the activity installations (e.g. accessibility, condition and restriction) were drawn from the BRAT-DO; the qualitative items about the accessibility, condition and restriction of water-sprinklers, skate parks and schoolyards were newly added. All thirteen new items were included to assess features of parks that would likely appeal to or be relevant specifically to youth physical activity in parks. These items include the presence of schoolyards, skate parks, water sprinklers, and qualitative general impression items such as overall safety and appeal for youth.

Further, in a later section it says that 20 items were drawn from the post, 17 of which could be directly compared. How does this relate to the previous numbers?

Thank you for pointing out this discrepancy, this has been corrected, line 304.

b. Can you provide a little more context of some of the scale items (what does “least positive” mean? What type of scales are the 4pt and 5pt scales – rating scales, ore more like the pool length?)

The following has been added to the section on the scale items, lines 193-200: “The response options on the tool include binary yes/no, or present/absent responses (n=61), 3-point scale items (1 being the most favourable response and 3 being the least favourable response, e.g. for presence of graffiti: 1=none, 2=some, and 3=a lot; except for the pool length item which had three options for pool length) (n=24), 4-point scale items, such as the cleanliness of water sprinklers where 1=very clean, 2=clean enough, 3=not at all clean, and 4=impossible to evaluate (n=2), 5-point scale items, such as water sprinklers condition where 1=no deterioration, 2=presence of deterioration without need for repairs, 3=significant deterioration requiring repairs, 4=under construction and 5=impossible to evaluate (n=3), and two text responses.”

c. Were the qualifying items regarding the activity installations added to the new tool, or were those removed?

These items, drawn from those on the BRAT-DO, were added to the PARK tool. This has been clarified in the text as the following: “For example, in the PARK tool, every activity installation item had three adjoining qualifying items: check if the installation is accessible, in good condition, and restricted.” Lines 205-207.

d. What are the implications of removing items from a tool that aren’t relevant to a particular study site? Doesn’t that limit the tools utility for different contexts (which is one of the objectives of this new tool?)

Perhaps the wording of this sentence was wrong, and it has been changed to: “The POST and BRAT-DO both contained a number of items regarding water features, including beachfront features, which were not applicable for the Montreal study site.” Lines 207-208.
In other words, rather than the items not being relevant, they were not applicable to our Montreal setting in much the same way there would not be questions about skating rinks in a tool developed for the Australian context. However, there are more general questions about the presence of an important body of water, as is written in the following sentence, lines 208-210: “These items were consolidated into one primary and one sub-item: Important body of water present, and if yes, are there sportive aquatic activities present.”

It may be that the more general question about an important body of water makes this item more generalizable to a variety of contexts than the more specific beachfront questions do.

e. Were the changes described at the end of this section changes as a result of the pilot test, or a modifications made prior to the pilot test? It isn’t clear from how it is worded.

Thank you for pointing out that this was not clear. It has been changed to the following, lines 216-220: “All item modification and development were conducted via an iterative process that was conducted at the beginning of the tool pilot stage (e.g. after a pilot run some items were modified if they caused confusion for observers) and the final items were agreed upon through discussion and consensus by the study team. The tool was piloted among observers from diverse ethnic backgrounds in their early twenties (n=12).”

f. I recommend breaking this long paragraph up into 2 or possibly 3 paragraphs.

This long paragraph is now broken up into 4 paragraphs.

4) I am concerned about the appropriateness of test-retest evaluations that span an average of 61 days, and up to 448 days. What if parks change over this period of time as they often do? How does that impact your analysis and evaluation?

We agree with the reviewer and are also concerned about the appropriateness of test-retest evaluations that span such a long time frame. That is why we directly address this issue in the first paragraph of the study limitations section, lines 469-475: “First, the study was not initially designed to assess test-retest reliability, resulting in a low number of test-retest occurrences and an inability to assess intra-rater reliability for all items. The wide range of days between tests was not controlled for and this may have compromised the validity of the test-retest results. The mean number of days between tests was 163, or approximately 5 months. This time lag likely resulted in non-differential misclassification thus underestimating the test-retest reliability, along with an increasing chance of substantive changes to park features between the first and second tests.” We hope the reviewer finds this satisfactory.

5) On page 13 it is important to explain how the POST data was obtained (from published data rather than collected data).

The following sentence has been added to this section: “Results are published and available on an institutional website [31].” The reference includes the URL where these results can be found. Lines 303-304.

Please also include a statement regarding why the BRAT-DO was not included (no corresponding item level data published for BRAT-DO)

The following sentence has been added to this section: “An item-by-item comparison between
items shared with the BRAT-DO could not be conducted because no corresponding item-level reliability estimates were published for the BRAT-DO.” Lines 307-309.

6) It would be helpful to provide more details on which items had low reliability (inter and intra) in the text results. Further, although readers are referred to the table, the tables only include the domains, and still do not provide any indication regarding the type of item that was low. In the discussion, several specific items were discussed, so reliability of those items could be provided in the results.

These details have now been added to the results section, lines 332-338: “A small number of the activity installation qualifying items (tennis restriction, basketball condition, track condition, and pool length) had <75% agreement, with the lowest percent agreement being 70.2% for pool length. Other items with low agreement include presence of shade, graffiti and litter (67.5, 69.3 and 67.3 % agreement respectively). The presence of pedestrian safety items had low percent agreement (73.8%) and the subjective general impression items, all had low agreement (ranging from 58.2% to 61.1%) except for attractive for bicycling (81.5%).”

As well as the following addition, lines 341-347: “Items for which there was low agreement include the activity installation condition items for tennis condition (kappa = 0.18), basketball condition (kappa = 0.25), trail condition (kappa = 0.26), 6-plus play area condition (kappa = 0.30), multi-use area condition (kappa = 0.24) and school yard condition (kappa = 0.17). The skate park restriction item also had poor agreement (kappa = 0.10), as did the presence of vandalism item (kappa = 0.22). Two of the general impression items had poor agreement: overall safe (kappa = 0.35) and overall attractive/pretty (kappa = 0.36).”

Also, is there any correspondence with the response type and reliability? For example, did binary items do better than 4 or 5pt scales?

There was no correspondence with the response type and reliability estimates. The only correspondence between the type of item and its reliability estimates were regarding the objectiveness of the item. Questions that were more subjective were less reliable than more objective items. This is addressed in detail in the discussion section, lines 402-436.

Further, was there any correspondence between items that were low intra-rater reliability and low inter-rater reliability? In other words, were there a few items that were low across multiple measures?

Yes, for seven items, there was poor reliability for percent agreement only, between both intra- and inter-rater reliability. The following details have been added to the manuscript, lines 360-368: “There were seven items that had both poor intra-rater reliability and poor inter-rater reliability based on % agreement. The presence of graffiti had poor intra- and inter-rater percent agreement (67.5% and 69.3%, respectively), as did the presence of litter (52.5% and 67.3%, respectively). Traffic calming measures had poor percent agreement for both intra- and inter-rater agreement (55% and 63.2%, respectively), as did pedestrian safety features (65% and 73.8%, respectively). Three general impression items, overall appealing for youth, overall safe, and attractive for walking had poor percent agreement for both intra-and inter-rater agreement (60% and 60.1%, respectively for overall appealing for youth; 70% and 60%, respectively for overall safe, and; 67.5% and 66.6%, respectively for attractive for walking).”

Were these the items included in the discussion?
A new paragraph has been included in the discussion about these items, lines 438-448: “A small number of items demonstrated poor percent agreement on both the intra- and inter-rater assessments including the presence of graffiti and litter, traffic calming measures and pedestrian safety features, and overall appealing for youth, overall safe and overall attractive for walking. These items may have had poor agreement for the intra-rater assessment because of substantive changes to the environment, e.g. litter and graffiti could have been removed and traffic calming and pedestrian safety features could have been installed. Whereas for the poor inter-rater agreement for these items, it may be difficult for raters to adequately assess or identify graffiti, litter and pedestrian safety and traffic calming features. Regarding the poor intra- and inter-rater reliability of the general impression items, substantive changes in the parks could again explain the poor intra-rater agreement between visits while the very subjective nature of these items may make achieving a high inter-rater agreement difficult.”

7) Pg. 15 Comparison of the PARK and POST:

a. Line 329, I would give the actual percent agreement for that PARK item with moderate agreement given it is only one item, or, given the actual range for moderate agreement rather than <75%.

The actual percent agreements have been added into the body of the manuscript as suggested by the reviewer, lines 373-376: “Percent agreement was not available for one item on the POST and 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.”

b. Regarding the Kappa, it is stated that many of the 7 not on the same range were in similar range. Were any vastly different? From the discussion it sounds like there were some, so include more details in the results.

None of the 7 items shared between the POST and the PARK were vastly different. For each of the seven items that were not in the same range, they only differed by one qualifying category and no more. For example, the kappa for drinking fountain present was 0.918, or almost perfect agreement, on the PARK tool and the kappa for this same item on the POST was 0.746, or substantial agreement; only one category lower. Another example is the presence of park benches where the PARK tool had an estimated kappa of 0.679, or substantial agreement, and the POST had 0.877, or almost perfect agreement for this same item. Again, as can be seen in Table 2, none of the items that were not in the exact same range were more than one category different.

As the reviewer pointed out, this seems to be unclear in results section. The following has been modified in the results section to try to make this clearer, lines 378-383: “Seven items were not in the exact same range, although all were similar in range (e.g. drinking fountains present had almost perfect agreement on the PARK tool and had substantial agreement on the POST, or sufficient lighting had moderate agreement on the PARK tool and substantial agreement on the POST) and none differed by more than one qualifying category. In other words, there were no items shared between the POST and the PARK tool that had vastly different reliability estimates.”

c. In the discussion, what do these similarities and differences mean?

We believe that these similarities mean that these items are generalizable across at least two very different geographic contexts and may be reliable in other contexts as well. Please see the
following responses below. As well, the changes to the discussion section of the manuscript regarding this point are detailed below.

In particular, what are the implications of those with consistent lower to moderate reliability?

Interestingly, none of the items shared between the POST and the PARK had low reliability. The only item that had slightly lower reliability across both tools was presence of litter, and this was still in the moderate agreement category, which is considered acceptable. One implication of this is that, perhaps the litter present item may never achieve very high agreement and that a kappa of >0.40 can be considered quite good for this item.

Might those also be items to examine for refinement?

Given the consistently high agreement between the items shared between the POST and the PARK tool, we think these are reliable items that could be used on a number of park direct-audit tools in a variety of contexts. Please see the additions made to the manuscript regarding this section, below.

With 7 of the 17 items falling in different ranges, what are the implications of that?

Because the 7 items were not vastly different, but only one category different, and all still demonstrated very acceptable agreement (e.g. mostly only differing between substantial and almost perfect agreement), that these items are reliable in a variety of contexts.

Are these items not cross-context reliable then? Given that this cross-context comparison is set up as one of the main justifications of this study, I think it is important to adequately discuss the findings.

On the contrary, we believe that these items are very much cross-context reliable given that they demonstrated very similar if not exactly the same reliability across different contexts.

We have elaborated on the points above in the discussion section of the manuscript with the following, lines 389-395: “Because of the very similar and acceptable agreement of the items shared between the POST and the PARK tool, it can be argued that these items are likely generalizable across at least two very different geographic contexts and may be reliable in other urban contexts as well. The only item shared between the two tools that had slightly lower reliability (moderate agreement) was ‘presence of litter’. One implication of this may be that this item may never achieve very high agreement and that a kappa of >0.40 can be considered quite good for this particular subject matter.”

8) Abstract: I would number the purposes of the study in the abstract as is done in the main text, as it is currently hard to follow that sentence in the abstract.

The purposes of the study in the abstract are now numbered, as is done in the main text, lines 35-39.

Please also add the total number of items in the methods section of the abstract to help give context to the numbers given.

The total number of items has now been added to the methods section of the abstract, line 43.
I would also argue that the 576 parks evaluated is part of the methods not the results.

The 576 parks evaluated has now been removed from the results section and added to the methods section of the abstract, line 44.