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: 2 Date: 16 May 2015

Author’s response to reviews: see over
Dear Dr. Latkin,

Please find enclosed our 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 and tables where requested. The point-by-point reply to both reviewers is enclosed with this letter.

We look forward to your editorial response.

Sincerely yours,

Madeleine Bird, M.Sc.
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 1 and 2 - Response

Reviewer 1

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.

The reviewer makes a very good point here, and indeed, this was the original intention of the study PI. To clarify, the original intent was to use an existing park direct-observation tool (e.g. the POST) to evaluate the parks around the homes of the youth in the QUALITY Cohort. However, it became clear that no single existing tool was able to meet the needs (notably in terms of efficiency, relevance, and reliability) of the study objectives. 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 current study. Given these changes, an independent reliability study is warranted; furthermore, this manuscript will be useful and necessary for replication and application of the tool in other studies.

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).

We agree that our description was incomplete; the sentence regarding the Kaczynski et al tool has now been removed.

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.

One of the purposes of table 1 is to present reliability results of different tools that have been evaluated in different geographic locations. In order to be more explicit about this, four new columns have been added to Table 1: 1) Test site, 2) Lowest Estimate, 3) Highest Estimate, 4) % Items ≥ 0.40 kappa or ≥ 70% agreement. These four new columns have been highlighted in Table 1.

In addition, a comment about reliability across the different tools tested in different geographic locations has been added to the body of the manuscript. The comment included is the following: Table 1 shows a comparison of the reliability estimates of
direct observation park evaluation tools. All but two of the tools were tested in the USA, while two were tested in Australia. All reported that the majority of items were at least moderately reliable or higher (i.e. more than 50% of items were $\geq 0.40$ kappa or $\geq 70\%$ agreement), which speaks to the general overall reliability of currently available direct-observation park audit tools. (Lines 97-102.)

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.
We agree and have revised the list to include those where the main emphasis is park audits. Thus, the SOPARC evaluation tool has been removed from Table 1 and from the reference list in paragraph 3 of the background (line 97).

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).
Many items ($n=22$) that appeared on both the POST and BRAT-DO were adapted for the present tool (e.g. presence of basketball courts). Although not identical, the qualitative items about the activity installations in the PARK tool (i.e. all the accessibility, condition and restriction items) are similar to items on the BRAT-DO. The majority of the environmental assessment, services, and safety items were drawn from the POST ($n=23$), and from the BRAT-DO ($n=21$) albeit with significant changes for several items. For example, the POST has a series of specific questions about shade provided by trees, whereas the PARK tool has a single item about the presence of shade. Eight new items were included to assess features of parks that would likely appeal or be relevant specifically to youth physical activity in parks but that were not present either in the POST or the BRAT-DO. These items include the presence of schoolyards, skate parks, water sprinklers, and qualitative general impression items such as overall safety and appeal for youth (lines 158-169).

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”?
Our language was unclear and the wording has been revised. Indeed, the new items were not drawn from any existing instrument but were newly developed for the PARK tool. This information has been corrected in the manuscript, lines 166 and 42.

What type of response options were used for each item (e.g., purely present/absent, or other scales)?
There were a number of different types of response options. The items that evaluated activity installations and areas were mostly evaluated with present/absent (or yes/no) response options. Exceptions include:
- Pool length was evaluated on a 3-point scale ($1 = \text{under } 25 \text{ m}; 2 = \geq 25 \text{ m}; 3 =$
impossible to evaluate);

- Pool and water sprinkler condition was evaluated on a 5-point scale (1 = no deterioration; 2 = presence of deterioration; 3 = significant deterioration requiring repairs; 4 = under construction; 5 = impossible to evaluate);

- Pool and water sprinkler cleanliness was evaluated on a 4-point scale (1 = no deterioration; 2 = presence of deterioration without need for repairs; 3 = significant deterioration requiring repairs; 4 = impossible to evaluate).

The following items were all evaluated on a 3-point scale (from 1 being the positive evaluation to 3 being the least positive evaluation):

- Shade,
- Graffiti,
- Vandalism,
- Litter/garbage,
- Garbage bins, Drinking fountains,
- Picnic tables,
- Benches,
- Bleachers,
- Public Toilets,
- Condition of Toilets,
- Change room,
- Condition of Change room,
- Parking,
- Streets adjacent to the park are local,
- Streets adjacent to the park have traffic calming measures,
- Streets adjacent to the park have pedestrian facilitation measures,
- Is the park attractive for youth,
- Is the park safe,
- Is the park pretty/attractive,
- Is the park appealing for walking,
- Is the park appealing for cycling,
- Is the park appealing for active play.

Two responses require a text response (specification of rental equipment if any is available to rent and specification of decorative or cultural physical elements).

Information about the response options has been added to the tool development section, as requested by the reviewer. These changes are highlighted on lines 169-173.

The audit tool has been provided in its entirety in additional file 1 for the reader to gain specific information about the response scale for each item.

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)?
During pilot testing, a number of changes were made to the POST and BRAT-DO items that were used on the PARK tool. Changes were made to items that had systematically poor agreement, that were less relevant to the study site, that were confusing to observers, or that required more detail. Specifically, every activity installation item had three additional qualifying items: check if the installation is accessible, in good condition, and restricted. The POST and BRAT-DO both contained a number of items regarding water features that were not relevant for the study site. 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. The POST and BRAT-DO also both had a number of items about dog related amenities (e.g. “Are dog litter bags provided?” on the POST and “Are there any signs specifying that dog owners must dispose of pet droppings?” on the BRAT-DO) were not included except for a modified item that asks the observer to check yes if dogs are not allowed in the park. Safety items from the POST, such as, “From the centre of the POS [public open space], how visible are surrounding road” were modified to “At least 1 street visible from the centre of the park” (lines 173-186).

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? A similar comparison between items was not made to previous BRAT-DO reliability estimates in other contexts because the reliability estimates from Bedimo-Rung et al. [1] were published as average percent agreement by conceptual domain and not item by item. Therefore, it was impossible to do an item-by-item comparison. Bedimo-Rung does publish the percent agreement of selected items on the BRAT-DO, however, there are only two items from this table that also appear on the PARK tool.

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). Yes, the reliability of only 17 items were compared between the PARK and the POST. This was stated explicitly in the abstract (line 45), and has been added to the methods section (line 280), in the discussion section (line 382), and in the limitations section (lines 436-438).

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.

The subheading has been changed to “Park identification and sampling” (line 197).

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.

The sampling plan information in the Results section has been moved to the methods section (lines 220-227).

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).

The definition of park used for this study, as well as the inclusion and exclusion criteria were the following: A park was defined as a public open space large enough to play a game of catch or roughly half the size of a soccer field (e.g. approximately 50 m long by 30 m wide). This included parks adjacent to schools and schoolyards. If the observers were not sure, they were instructed to look for the name of the park on a sign. Spaces that were exclusively passive, e.g. not large enough or equipped for physical activity (e.g. a concrete area with park benches only) were excluded, as were parks that were exclusively equipped with amenities and installations for children 5 years and under, and where there was a sign explicitly stating that the area was restricted for children 5 years and younger. Cemeteries and golf courses were also excluded.

This information has been added to the Park identification and sampling section (lines 203-211).

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?).

Observers were between the ages of twenty-one and thirty, seven female and two male, recruited through University employment services. Observers were mostly undergraduate or recent graduate Kinesiology or Community Design students.

This information has been added to the manuscript under the Observer training section (lines 229-232).

Were each of the nine observation training days full days of training? Yes, the nine observation training days were full days of training. The 9-day training on park audits was embedded in a larger neighbourhood environmental audit around the homes of the QUALITY Study participants. This information has been added to the manuscript (line 233-234).

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. This was indeed the case; a qualification regarding the training needed to correspond with the tool’s use has been added to the conclusion (lines 447-448).

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?

Each park was audited by two observers who evaluated the parks independently but at the same time on the same day of the week. Observer pairs were assigned on the morning of
every day of observation, so that the pairs were not always the same two individuals.

This information has been added to the Reliability assessment of observers’ section (lines 255-257).

In addition, given some confusion about the contents of this section, the title of this section has been revised to “Observer reliability and validity assessment” (line 254).

(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. The intra-rater reliability section was expanded and the test-retest methods used were explained.

The changes made to the manuscript are the following (lines 258-261):

Intra-rater reliability was assessed using test-retest methods wherein the same observer audited a park on two separate occasions, and the results of the independent audits of the same park by the same observer on different occasions were compared.

Also, provide the details included in later sections regarding the time span between the assessments as this is relevant methods information.

This information has been moved to this section, lines 261-262.

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.

The reason it was not possible to assess intra-rater reliability for all the items on the tool was originally explained in the results section. The explanation has been expanded and clarified and moved to the methods section (lines 264-268). This is mentioned again briefly in the results section (lines 316-318).

The explanation provided in the methods section is the following:

Reliability estimates of a test-retest can only be calculated using complete data with response item variation. For example, if there is no tennis court in any of the parks that were re-audited by the same observer, then the intra-rater reliability of this and associated items (e.g. tennis court condition) cannot be evaluated.

(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 section has been completely revised because it was, as the reviewer kindly mentioned, unclear (lines 270-277). During a one-week period in June 2008 a validity assessment of the observer’s responses took place in which the same park was evaluated
by the observer trainer, and all the observers at different times over the week. The observers were unaware that the validity assessment was taking place. The responses from the observers were then compared to the responses of the trainer, which were considered the gold standard.

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?

Only one park was assessed by all the observers over the weeklong period, and observer responses were compared with the responses of the trainer. The trainer responses were considered the gold standard (see lines 270-277).

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?)

Results from the observer responses’ validity assessment (observer responses compared to the gold standard on an example park) were moved from the methods section to the results section (lines 311-313).

Were park audits conducted as part of the training also included as part of the sample?)

If a park audited for training purposes was also in the study sample, the park was re-evaluated for the study. This information was added to the manuscript, lines 275-277.

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?

We agree with the reviewer. The methods of objectives 2 are missing and have been added to the manuscript, and results of objective 3 have also been added (please see below).

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.

A section titled “Comparison of reliability estimates between POST and PARK tools” has been added (line 278) and a description of the methods for this comparison has been added. The methods description added is the following (lines 279-285):

Twenty items on the PARK Tool were drawn directly from the POST, seventeen of which could be directly compared with items assessed in Montreal, Canada. Inter-rater reliability of the POST, previously assessed in Perth, Australia, was estimated by calculating Cohen’s kappa and percent agreement between raters. Comparison of reliability estimates were done visually and qualitatively, using the percent agreement cut-offs according to Saelens and colleagues [18] and the kappa interpretation guidelines suggested by Landis and Koch [31], both described in the next section.

The methods also needs more information the POST data (number of parks, where, when).

Reliability of the POST was assessed in 2005 among 516 parks in Perth, Australia (line
Also, more is needed in the results regarding this (in addition to referring to table 2, describe some of the findings in the text).

Details of the findings have been added to the text:
Inter-rater reliability estimates for 17 items were compared and found to be of a similar magnitude (see Table 2). Fourteen of the seventeen items compared were $\geq 75\%$ agreement on both the POST and the PARK tool. 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 ($<75\%$ agreement) while the POST had good to excellent agreement ($\geq 75\%$) for these same items. The kappa coefficients fell in the same ranges for 10 of the seventeen items (4 items had kappa $>0.80$, 3 items had kappa $>0.60-0.80$, and 3 items had kappa $>0.40-0.60$ for both items on both tools). Seven items were not in the same range, although many were similar in range (e.g. drinking fountains present had almost perfect agreement on the PARK and had substantial agreement on the POST, or sufficient lighting had moderate agreement on the PARK and sufficient agreement on the POST) (lines 324-334).

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. We agree with the author and her concerns about objective 3. Therefore, we have modified the wording around objective 3: To compare the identification of the existence of parks through direct observation and desktop park identification (lines 38-39 in the abstract and 133-135 in the background section), and we have made the methods for objective 3 more explicit (line 218-220). In addition, we have included a results section specifically for objective 3 (lines 335-338).

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?
Summary tables of the information in additional file 2 and 3 have been created and added to the main paper. The tables are the inter- and intra-rater reliability results by tool domain for Cohen’s kappa and percent agreement. The number of items in each domain, the average kappa and agreement, the highest kappa and agreement, the lowest kappa and agreement, and the number of items $\geq$ either .40 for kappa or 70% for percent agreement as well as the percentage of items $\geq$ either .40 for kappa or 70% for percent agreement are shown in tables 3-6. Table 3 is referred to on line 305 and Table 4 on line 310. Table 5 is referred to on line 320 and Table 6 on line 321. Additional (much more detailed) files 2 and 3 remain available to consult should the reader be interested.

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?
More details were added to this section. The details added are the following:

Modification to the Overall safe item could include a link to the objective safety items already checked in the tool. Thus, to rate the park as ‘very safe’, the observer would have had to have checked ‘yes’ to a minimum of two of the following safety items: sufficient lighting; at least 1 street visible from the center of the park; at least 1 house visible from the center of the park. To reply that the park is ‘safe enough’, the observer would have had to have checked ‘yes’ to one of the above listed safety items. If the park is rated as ‘not safe’, then none of the above safety items would have been present. For the Overall attractive/pretty item, an objective benchmark could be that there is a decorative piece present that is not in disrepair (such as a sculpture/statue or a fountain).

These additions are included on lines 349-357.

Other additions are on lines 360-361: One way this item may be modified is to add objective qualifiers such as presence of broken or damaged installations.

Also, the additions listed above are referred to on line 372 in the discussion.

What would enhanced observer training include (beyond the already 9 day training)? Enhanced observer training would include requiring observers to successfully pass an observation reliability test on test sites prior to data collection. For example, if observers are unable to achieve an overall reliability of a 0.60 kappa or 70% agreement after training, then they are not retained for the study evaluation. (lines 374-377).

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. We hope the reviewer is satisfied with the additions made to the methods and results section concerning objectives 2 and three, as detailed above.

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. These concerns have been addressed, detailed above. We hope these are satisfactory to the reviewer.

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: http://www.sciencedirect.com/science/article/pii/S0143622812001361
We understand the concerns regarding objective 3. We hope that the concerns have been addressed (please see above), through modifying the wording around objective 3,
explicitly adding a statement in the methods section regarding this objective (lines 218-220), and providing a results section specifically for this objective (lines 335-338).

We thank the reviewer for suggesting this excellent article, and the article is now referenced in the discussion section of this paper (lines 398-404).

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. 
Wording regarding the lack of established validity of the tool has been added to the conclusion (line 450).

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.
The importance of the training has been added to the conclusion, line 447-448.

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.
Additional recommendations for future research have been added to the conclusion. These additions are found on lines 452-453.

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.
The purpose of the study, as stated in the abstract and the main text, has been edited so that it is the same in both places (lines 34-39 and lines 130-135).

2) First paragraph in background: include a citation for the remainder of the 2nd sentence regarding the adult morbidities.
Two citations were added (numbers 4 and 5) regarding obesity and type-2 diabetes and obesity and cardiovascular disease. (Line 71.)

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.).
Examples are provided regarding what is meant by park improvements: The park improvements include fitness zones appropriate for all fitness levels and individuals 13 years or older, renovations to soccer fields (e.g. artificial turf to replace dirt fields and lighting), and installation of an all-abilities playground and landscaping. This information has been added to the manuscript (lines 83-86).

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.
Citation numbers have been included for each reference in Table 1.

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. The citation for the POST has been changed to Giles-Corti et al., 2005 [2]. This study includes more background on the tool and the source of information in Table 1. This change is reflected in Table 1, as well as in the references in the third paragraph of the Background section (lines 97 and 102).

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? The idea we wished to convey was that there should be more reliability assessments, not that existing assessments are unsatisfactory. The wording has been changed in the text, line 124-128.

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. The sentence regarding the audits being embedded in a larger study was moved up to the previous section (lines 225-227).

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…” The sentence has been changed to read, “… Canadian youth lag behind other nations …”. Line 68.

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. This section was removed from the manuscript, as suggested by the reviewer. (Section removed between the sentences on line 143.)
Reviewer 2 – Response

Minor Essential Revisions

There are only two small issues that were unclear to me:

page 8, line 171: The tool underwent expert consensus and was piloted among youth from diverse ethnic backgrounds ....

What does this mean? It would be helpful to the reader to clarify the intent behind and what exactly the expert consensus consisted of, as well as what the piloting process resembled.
The section on pilot testing has been expanded greatly (lines 173-189). The sentence in question has been changed to the following: All item modifications were agreed upon through discussion and consensus by the study team, and the tool was piloted among observers from diverse ethnic backgrounds in their early twenties (n=12).

page 10, line 223: the development of the tool began in 2008, and it's just being published in 2015? I understand there were hundreds of parks that were visited in the development of the tool, and that this may have contributed to the time lag in test-retest results, as indicated in the limitations section. I think it would be helpful to readers if the authors comment on the lengthy nature of this study, and if it influences their tool at all.
We understand the reviewer’s concern and hope we have addressed the time intensive nature of direct observation in the study’s discussion section (lines 404-405). Although there is a time lag between the beginning of the study and the date of the present manuscript, we do not believe that this has an impact on the validity of the reliability estimates presented in the manuscript.

REFERENCES