PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Choosing appropriate tools and referral criteria for vision screening of 4- and 5-year-old children in Canada: a quantitative analysis.
AUTHORS	Nishimura, Mayu; Wong, Agnes; Cohen, Ashley; Thorpe, Kevin; Maurer, Daphne

VERSION 1 – REVIEW

REVIEWER	Bruce Moore OD New England College of Optometry Boston USA I do some very limited consulting to Welch Allyn on the Spot device. No other financial disclosures.
REVIEW RETURNED	16-Jun-2019

GENERAL COMMENTS	This is a very well designed study and is extremely well written. Very nice job. It is an important contribution to the literature and definitely deserves publication. The authors understand the issues underlying vision screening of young children. They are attempting to update the VIP results with the current generation of technology. They should point out that the PVS is now called the Blinq and they should indicate which software version each of the photoscreeners had installed when used in the study, as there are updates since then. There are several issues and clarifications that I think ought to be addressed in a final revision, as indicated in my comments below. In general, this is an excellent paper and one that will be widely read and appreciated by the community. This is a very well designed study and is extremely well written. Very nice job. It is an important contribution to the literature and definitely deserves publication. The authors understand the issues underlying vision screening of young children. They are attempting to update the VIP results with the current generation of technology. They should point out that the PVS is now called the Blinq and they should indicate which software version each of the photoscreeners had installed when used in the study, as there are updates since then. There are several issues and clarifications that I think ought to be addressed in a final revision, as indicated in my comments below. In general, this is an excellent paper and one that will be widely read and appreciated by the community. Small point: the Spot and PlusOptix are more properly described not as autorefractors but as photoscreeners entailing different design considerations and purposes than autorefractors. The Randot Preschool Stereoacuity Test used in the screening phase is far more sensitive (and frankly better in every way) than

	the Titmus stereo test used in the gold standard exams, the reverse of typical comparisons between screening and gold standard. The authors probably should explain or comment on this point. Actually same comment for the PVS versus the cover test for strabismus and eye position. Page 9 line 26: The photoscreener in reference 52 was from Atkinson and was not even remotely the production model of the ones in the study, and it was built to very different standards and design. I think the authors somehow got this reference mixed up with some other study. Reading the Atkinson paper referenced, they did not use a PlusOptix, but they did use as they described it, a VPR-1 videorefractor (Clement Clarke International Ltd.). This needs to be cleaned up. I am not aware of any photorefractor paper with a sensitivity result of 99% with a specificity of 82%, but perhaps there is. The authors did mention that both Spot and PlusOptix have been widely reported to have deficiencies in detecting moderate to high hyperopia, obviously among the most important clinical associations with amblyopia. In many of the previous papers on photoscreening, significant astigmatism, which is of course often associated with significant hyperopia, does get readily flagged by both devices, but they often do not detect the associated hyperopia in those individuals. It can be argued that “who cares how those kids get detected”, but it leaves those with hyperopia only being missed with frequency. This might be discussed somewhere in the paper. The data presented on the sensitivity and specificity of the Spot and PlusOptix for sphere in Supplementary Table 1 in fact make this point very clearly. If one looks at the PlusOptix subjects with +2.00D or greater sphere, sensitivity drops to below 50% with specificity of 93% or greater, and with the Spot, subjects with +1.50D or greater hyperopia drop below 50% sensitivity with specificity of 99%. For cylinder on both devices -2.00D (or so) is the inflection point where you get to about 50% sensitivity with greater than 90% specificity. Bottom line, both devices are missing lots of kids with hyperopia but catching those with astigmatism. I think this point needs to be discussed as something quite important. In Table 8, the AUC for spherical equivalent of both the Spot and PlusOptix are .85 or better, seemingly indicating very good results. There is a disconnect here somewhere. The comment on page 15 line 3 touches on this point of the photoscreeners catching astigmatism preferentially and the low AUC for sphere for both devices. Another point that I noticed is the overall incidence of the disorders in Table 3. Amblyopia is 5.9% (3.4% unilateral, 2.5% bilateral). Astigmatism 14.2%, while hyperopia only 5.4%. Given the high % of astigmatism and amblyopia, especially bilateral, one suspects that the incidence of significant hyperopia seems low. Why?? The 5.4% is also much below that found in the MEPEDS and BEPEDS studies. Section 2.5 what were the technique / details of the VA testing? The AUC figure is apparently garbled visually in the pdf. Is this a problem in only my copy or is it in need of reformatting?

	Page 13 line 42. The specificity of 64% for the 3 tests of VA, stereo, and refraction is quite low and will result in a substantial numbers of over-referrals. VIP used the 90% specificity to reduce the number of those over-referrals to a manageable level. This is critically important for any public health system where the availability of pediatric vision care and patient access to that care is frequently going to be limited, especially in rural areas, where overwhelming limited availability for access to care with too many normal kids will have potentially dramatic repercussions for those with real vision problems. In fact, in the VIP paper that analyzed combinations of tests versus single tests, that issue was central to the discussion and an important concern, then and now, for the health care system. The comment on Page 13 line 45 of the rationale for using the 3 instead of 4 tests was along those lines, but again, 64% specificity results in a whole lot of over-referrals. On Page 14, the discussion about the Spot versus the PlusOptix is along these lines, making clear the excessive number of over-referrals with the PlusOptix compared to the Spot, but still significant for the Spot. Furthermore, increasing the number of screening tests performed on each child increases the time required for screening each child, resulting in a decrease of "throughput" of kids screened in an actual school setting, a potential decrease in cost effectiveness, and the possibility that more tests will cause fatigue in some kids leading to more tests results becoming "untestable", thus further increasing over-referrals. Bottom line is more tests will likely pick up more kids but at the cost of increasing over-referrals. It is easy to imagine this being especially so in younger kids with the Randot Preschool Stereoacuity Test, which is longer and more complex than alternative forced choice stereo tests (as used in VIP). The issue of critical line versus threshold testing was not apparently discussed in this paper. These were all important considerations in the design and the results from the VIP Study. The authors might consider these points. Page 16 line 28. A point about the comment of VIP specificity set at 94%. That was calculated just to compare to the "preferred" specificity of 90%, and was done to show how sensitivity would be affected by even that modest increase in calculated specificity. Table 1. Interesting that the referral criteria for both photoscreeners were set the same as the AAPOS screening guidelines. I thought that there are differences between the devices based on different software and considerations by the manufacturers. Was this modified from the manufacturer-established software by the authors?

REVIEWER	Professor Jan Atkinson University College London , London , UK
REVIEW RETURNED	29-Jul-2019

GENERAL COMMENTS	This is a well- planned and thoroughly reported study on an important question in clinical vision screening in young children. Its value is due to a) the screening is carried out with a relatively large unselected group of typically developing children, rather than being based on a group with suspected visual problems, which is often the case in previous studies based in private ophthalmological clinics (perhaps a little more emphasis of this point should be made in the manuscript in the introduction and

	conclusions); (b) the care in which it analyses the comparative advantage of different screening tests and their combinations. It is also insightful in pointing out that childhood vision screening carried out in different healthcare contexts will have different priorities for maximising sensitivity vs minimising false referrals. It also includes some means of testing (Plusoptix autorefractor, PVS) that were not available in the earlier studies by the Vision in Preschoolers group. For these reasons the work deserves to become widely known among those responsible for young children's eye care. Some points which could benefit in revision: 1. The essential criterion for amblyopia is correctly stated as a deficit in best corrected acuity. However the description of the 'gold standard' eye exam does not make clear if and when acuity was tested wearing such a correction – the full benefit of the spectacles may not become apparent for a short time as the child adopts appropriate accommodation ('relaxing into glasses'). 2. p.11 refers to the prevalence of eye problems but it's not quite explicit as to what constituted a 'problem' If these are the criteria listed in Table 2 then this should be stated. In addition, the breakdown in Table 3 is a little confusing – presumably all the 6 conditions listed under 'risk factors' contributed to the total of 170 cases, but the total of the conditions is greater than this – the legend should make clearer that some children had more than one of these risk factors, as well as the overlap between risk factors and overt amblyopia. 3. The figure of 26.5% of children with confirmed vision problems seems high – I have not checked in detail all the references cited as refs 12-22, but for instance the Baltimore study cited as ref 14 found 2-3% with strabismus (3.9% here) and 0.8-2.5% with amblyopia (5.9% here). Some comment would be helpful – does this reflect in any way the likely low SES in the largely recent immigrant population which appears to dominate the present sample? Alternatively, the high incidence of significant astigmatism reported in the present study is reminiscent of results from a study on a Native American population (Harvey, Dobson et al, 2011) and other publications where differences in refraction have been reported across different racial groups. 4. A minor point is the discussion of alternative acuity tests for this age group. The authors point out correctly that the use of a 'crowded' acuity test is important, and cite, as alternatives to the Cambridge Crowding Cards (used in this study) the LEA and HOTV tests. Both the HOTV and LEA tests measure the extent of crowding by using a linear array. This can present a problem to young children of keeping their place on a line, and will result in reduced crowding for the end letter (or end picture) on each line of the chart. These problems are overcome in the Cambridge Crowding Cards in which each test card shows a single central letter, surrounded by crowding letters on all sides, with the child only having to understand what letter is in the 'middle' or 'centre' of the array and point to that letter on the matching board. Perhaps some comment on these advantages may be an appropriate addition to the paper.

VERSION 1 – AUTHOR RESPONSE

Reviewer #1 Comments

This is a very well designed study and is extremely well written. Very nice job. It is an important contribution to the literature and definitely deserves publication. The authors understand the issues underlying vision screening of young children. They are attempting to update the VIP results with the current generation of technology. They should point out that the PVS is now called the Blinq and they should indicate which software version each of the photoscreeners had installed when used in the study, as there are updates since then. There are several issues and clarifications that I think ought to be addressed in a final revision, as indicated in my comments below. In general, this is an excellent paper and one that will be widely read and appreciated by the community.

- We have added information about software versions of the photoscreeners (page 9) and the reference to Blinq. for the PVS (page 10). We also contacted the manufacturer and verified that it would be appropriate to refer to the device as “PVS” in this manuscript as we used a slightly older version of what is now called blinq.

Small point: the Spot and PlusOptix are more properly described not as autorefractors but as photoscreeners entailing different design considerations and purposes than autorefractors.

- We have switched the terminology to use “photoscreeners” instead of “autorefractors” throughout the manuscript.

The Randot Preschool Stereoacuity Test used in the screening phase is far more sensitive (and frankly better in every way) than the Titmus stereo test used in the gold standard exams, the reverse of typical comparisons between screening and gold standard. The authors probably should explain or comment on this point. Actually same comment for the PVS versus the cover test for strabismus and eye position.

- We agree with the reviewer that the Randot Preschool Stereoacuity Test is a more sensitive test for testing 4- and 5-year-olds; however, at the request of the optometrists we had available both the Randot and the Titmus Stereo test. We have clarified in the manuscript that both options were available (Section 2.4, page 11). All optometrists preferred using the Titmus test.
- The gold standard for ocular alignment is the cover test, so that was the protocol for the optometry exams. At the start of our study (2014) the PVS was not yet FDA approved and thus considered an experimental screening device; hence it was used for screening but not during the gold standard eye examinations.

Page 9 line 26: The photoscreener in reference 52 was from Atkinson and was not even remotely the production model of the ones in the study, and it was built to very different standards and design. I think the authors somehow got this reference mixed up with some other study. Reading the Atkinson paper referenced, they did not use a PlusOptix, but they did use as they described it, a VPR-1 videorefractor (Clement Clarke International Ltc.). This needs to be cleaned up. I am not aware of any photorefractor paper with a sensitivity result of 99% with a specificity of 82%, but perhaps there is.

- We apologize for the error and the correct reference for the statement “Plusoptix S04 showed a sensitivity of 99% and specificity of 92%” was reference 54, Matta et al. (2010). That correction has been made. There were additional referencing errors in that paragraph that have now been corrected.

The data presented on the sensitivity and specificity of the Spot and PlusOptix for sphere in Supplementary Table 1 in fact make this point very clearly. If one looks at the PlusOptix subjects with +2.00D or greater sphere, sensitivity drops to below 50% with specificity of 93% or greater, and with the Spot, subjects with +1.50D or greater hyperopia drop below 50% sensitivity with specificity of 99%. For cylinder on both devices -2.00D (or so) is the inflection point where you get to about 50% sensitivity with greater than 90% specificity. Bottom line, both devices are missing lots of kids with hyperopia but catching those with astigmatism. I think this point needs to be discussed as something

quite important. In Table 8, the AUC for spherical equivalent of both the Spot and PlusOptix are .85 or better, seemingly indicating very good results. There is a disconnect here somewhere. The comment on page 15 line 3 touches on this point of the photoscreeners catching astigmatism preferentially and the low AUC for sphere for both devices.

- Table 8 shows the AUC for spherical equivalent from Plusoptix and Spot in diagnosing hyperopia to be 0.856 and 0.886, respectively. These data show that using the optimal referral threshold, the photoscreeners can, fairly accurately, classify a child as having or not having hyperopia. How many children will be missed (i.e., rate of false negatives) will depend on the referral threshold chosen. As the reviewer correctly points out, choosing higher thresholds will lead to more children being missed. We have emphasized this point more clearly in the Discussion of the manuscript (see below), however, as Reviewer #2 has indicated, we believe that the strength of our manuscript lies in the summary of how sensitivity and specificity changes with different cut-offs so that policy makers can make informed decisions; not to recommend specific screening criteria.
- Discussion, page 16: "The optimal set of screening tools will depend on available resources as well as the social and political climate of a given jurisdiction. Adding more tests will result in fewer missed problems (higher sensitivity) but more over-referrals (lower specificity), which can overburden clinicians unnecessarily where access to eye care may be difficult. For example, the VIP study in the US examined how sensitivity of the screening tests changed with specificity targeted at 90%[19,39] and 94%[64] to prevent unnecessary referrals. In our data, both the Plusoptix and Spot photoscreeners had high specificity, above 90% for SK children and reasonable sensitivity (around 70%). This finding, in part, may be because our sample had a high prevalence of astigmatism, which photoscreeners detect more accurately than they do hyperopia or myopia.[40,47,53]"

Another point that I noticed is the overall incidence of the disorders in Table 3. Amblyopia is 5.9% (3.4% unilateral, 2.5% bilateral). Astigmatism 14.2%, while hyperopia only 5.4%. Given the high % of astigmatism and amblyopia, especially bilateral, one suspects that the incidence of significant hyperopia seems low. Why?? The 5.4% is also much below that found in the MEPEDS and BEPEDS studies.

- The surprisingly low rate of hyperopia may be explained by how we defined hyperopia: SE > +3.50 according to AAPOS 2013 guidelines for refractive errors that pose significant risk for developing amblyopia. MEPEDS (2013; Reference 21) reported a prevalence of 25.65% for hyperopia among Non-Hispanic White preschool children and 13.47% for Asian children, but the definition for hyperopia was SE >= +2.00D. BEPEDS (2009; Reference 22) reported a prevalence of hyperopia of 8.9% in White preschool children and 4.4% in African-American children, with hyperopia defined as SE >= +3.00D. The issue of varying definitions of hyperopia by different research groups and clinicians is beyond the scope of our paper; however, we have clarified that our prevalence estimates of visual problems are based on the definitions outlined in Table 2 (page 12, Section 3: Results, 1st sentence). We have also clarified that although we used AAPOS guidelines for defining visual problems for the purpose of research, the optometrists used their own clinical judgements about whether a child should be treated with spectacles (Section 2.5, page 11). "Optometrists made clinical judgments about whether a child should receive treatment (e.g., glasses) independently from these research definitions. Prescribing practices vary among clinicians, and thus "number of glasses prescribed" was not an *a priori* variable of interest in our study."

Section 2.4 what were the technique / details of the VA testing?

- We have clarified in Section 2.4 (page 11) that visual acuity testing at follow-up optometry exams was conducted for both near and far acuity. Screening was only at far (page 9).

The AUC figure is apparently garbled visually in the pdf. Is this a problem in only my copy or is it in need of reformatting?

- The PDF file we downloaded from the ScholarOne Manuscript website had no issues, so we are not sure of what happened. We will work with the editorial office to ensure that the final version is correct.

Page 13 line 42. The specificity of 64% for the 3 tests of VA, stereo, and refraction is quite low and will result in a substantial numbers of over-referrals. VIP used the 90% specificity to reduce the number of those over-referrals to a manageable level. This is critically important for any public health system where the availability of pediatric vision care and patient access to that care is frequently going to be limited, especially in rural areas, where overwhelming limited availability for access to care with too many normal kids will have potentially dramatic repercussions for those with real vision problems. In fact, in the VIP paper that analyzed combinations of tests versus single tests, that issue was central to the discussion and an important concern, then and now, for the health care system. The comment on Page 13 line 45 of the rationale for using the 3 instead of 4 tests was along those lines, but again, 64% specificity results in a whole lot of over-referrals.

- We agree with the reviewer that decision makers in many health care systems will worry about over-referrals. We have emphasized in the Discussion the point that with additional tests, sensitivity is increased at the cost of lowering specificity (i.e., more children without a visual problem will be referred – page 16, 2nd paragraph). As stated above, we made an effort to provide an objective summary of the data rather than to make specific recommendations because different communities will have different attitudes about over-referrals. For example, in Ontario, Canada, where the provincial healthcare system pays for every child (age 0-19 years) to have an optometry exam annually, optometrists advocate for 100% of children to have annual optometry exams. In such a climate, low sensitivity is more worrisome than low specificity.

On Page 14, the discussion about the Spot versus the PlusOptix is along these lines, making clear the excessive number of over-referrals with the PlusOptix compared to the Spot, but still significant for the Spot. Furthermore, increasing the number of screening tests performed on each child increases the time required for screening each child, resulting in a decrease of “throughput” of kids screened in an actual school setting, a potential decrease in cost effectiveness, and the possibility that more tests will cause fatigue in some kids leading to more tests results becoming “untestable”, thus further increasing over-referrals. Bottom line is more tests will likely pick up more kids but at the cost of increasing over-referrals. It is easy to imagine this being especially so in younger kids with the Randot Preschool Stereoacuity Test, which is longer and more complex than alternative forced choice stereo tests (as used in VIP). The issue of critical line versus threshold testing was not apparently discussed in this paper. These were all important considerations in the design and the results from the VIP Study. The authors might consider these points

- We have added a paragraph (page 17) in the Discussion to address the issue of “critical line versus threshold testing” as well as a discussion on whether fatigue may have impacted the results. “In pilot work, we found that most children failed the two behavioural tests (visual acuity and stereoacuity) if we started testing at the referral threshold. Thus, we chose to begin with the easiest levels (largest letters of 6/60 for acuity testing and largest discrepancy of 800 arcsec for stereoacuity) and to work towards more difficult levels until children made mistakes. We believe that this is an important strategy for school-based vision screening so that the screeners have time to build rapport with the children and to engage the child with the “games”. Although only two of the five screening tests required a verbal response, it is possible that some children failed more from fatigue than an actual visual problem as they went through the later tests. The order of tests was counterbalanced and there was no systematic effect of order on referral rate for any test, with referral rates ranging from 2% (children who completed Randot second) to 8% (children who completed acuity as their 1st test and as their 4th test). However, fatigue may account for the lower accuracy of the screening tools in JK compared to SK children, a result suggesting that fewer tests may be advantageous for that age group and not just from a cost-savings perspective.”
- As described above, we have added a sentence in the Discussion (page 16) to emphasize the point that more tests will lead to over-referrals.
- Below is a table with the referral rates on each test as a function of the order of the test (1 = 1st test for that child).

Order of Test	1	2	3	4	5
PVS	5.7%	6.6%	5.3%	4.4%	5.3%
Plusoptix	3.9%	4.8%	4.9%	5.9%	5.5%
Spot	4.9%	3.1%	6.2%	3.8%	4.1%
Acuity	8.0%	7.1%	6.3%	7.6%	6.6%
Randot	2.7%	2.0%	3.5%	3.1%	4.1%

Page 16 line 28. A point about the comment of VIP specificity set at 94%. That was calculated just to compare to the "preferred" specificity of 90%, and was done to show how sensitivity would be affected by even that modest increase in calculated specificity.

- We have modified our wording to more accurately reflect the purpose of the VIP studies referenced. "For example, the VIP study in the US examined how sensitivity of the screening tests changed with specificity targeted at 90%[19,39] and 94%[64] to prevent unnecessary referrals" (page 16).

Table 1. Interesting that the referral criteria for both photoscreeners were set the same as the AAPOS screening guidelines. I thought that there are differences between the devices based on different software and considerations by the manufacturers. Was this modified from the manufacturer-established software by the authors?

- The Spot (Welch Allyn) allows the user to input the referral criteria for hyperopia, myopia, astigmatism, and anisometropia, whereas the Plusoptix S12 comes pre-programmed with 5 options along a ROC curve (to maximize sensitivity or specificity), none of which matched the AAPOS guidelines for preschool children. Thus, all screeners were trained to examine the values of Sphere, Cylinder, and Spherical Equivalent, and to make the pass/refer decision based on AAPOS guidelines. We have added this point in the Methods (section 2.3, page 10). "The pre-programmed referral thresholds on the Spot and Plusoptix did not match these AAPOS guidelines and thus screeners were trained to manually make the decision of pass/refer based on the refractive error values."

Reviewer #2 Comments

This is a well- planned and thoroughly reported study on an important question in clinical vision screening in young children. Its value is due to a) the screening is carried out with a relatively large unselected group of typically developing children, rather than being based on a group with suspected visual problems, which is often the case in previous studies based in private ophthalmological clinics (perhaps a little more emphasis of this point should be made in the manuscript in the introduction and conclusions); (b) the care in which it analyses the comparative advantage of different screening tests and their combinations. It is also insightful in pointing out that childhood vision screening carried out in different healthcare contexts will have different priorities for maximising sensitivity vs minimising false referrals. It also includes some means of testing (Plusoptix autorefractor, PVS) that were not available in the earlier studies by the Vision in Preschoolers group. For these reasons the work deserves to become widely known among those responsible for young children's eye care.

- We have emphasized that our sample is of typically developing children in the “Strengths & Limitations” section, as well as in the Introduction & Discussion.

Some points which could benefit in revision:
1. The essential criterion for amblyopia is correctly stated as a deficit in best corrected acuity. However the description of the ‘gold standard’ eye exam does not make clear if and when acuity was tested wearing such a correction – the full benefit of the spectacles may not become apparent for a short time as the child adopts appropriate accommodation (‘relaxing into glasses’).

- We have included a paragraph in the Discussion that addresses this concern – that amblyopia may be resolved fairly quickly after wearing spectacles. We now acknowledge this fact as a limitation of the study and that we may have overestimated the prevalence of amblyopia in our sample (page 18):

  “One limitation of our study is that we may have over-estimated the prevalence of amblyopia (5.9%) because it sometimes resolves quickly after spectacle correction. Although acuity was re-measured at the time the child received the glasses, follow-up assessments were not made. Had they been possible, the prevalence of amblyopia might have been lower. Nevertheless, those cases would still have been included in the count of children identified as having eye problems requiring treatment.”

2. p.11 refers to the prevalence of eye problems but it’s not quite explicit as to what constituted a ‘problem.’ If these are the criteria listed in Table 2 then this should be stated. In addition, the breakdown in Table 3 is a little confusing – presumably all the 6 conditions listed under ‘risk factors’ contributed to the total of 170 cases, but the total of the conditions is greater than this – the legend should make clearer that some children had more than one of these risk factors, as well as the overlap between risk factors and overt amblyopia.

- We have explicitly stated that the problems are those described in Table 2. (Section 2.5, page 11)

3. The figure of 26.5% of children with confirmed vision problems seems high – I have not checked in detail all the references cited as refs 12-22, but for instance the Baltimore study cited as ref 14 found 2-3% with strabismus (3.9% here) and 0.8-2.5% with amblyopia (5.9% here). Some comment would be helpful – does this reflect in any way the likely low SES in the largely recent immigrant population which appears to dominate the present sample? Alternatively, the high incidence of significant astigmatism reported in the present study is reminiscent of results from a study on a Native American population (Harvey, Dobson et al, 2011) and other publications where differences in refraction have been reported across different racial groups.

- We have acknowledged the limitation in our study that we did not collect demographic data from our sample and thus cannot explain why there is a higher prevalence of visual problems than in other studies. One possible explanation is the high number of recent immigrant families at the school (page 18): “Another limitation is the lack of demographic data that might help explain the high percentage of screened children (26.5%) found to have a visual problem. The school serves an immigrant community and it is known that race and ethnicity affect prevalence rates of refractive errors. For example, astigmatism is more common among

Native American, African American, and Hispanic children compared to non-Hispanic White children. [65-66]"

4. A minor point is the discussion of alternative acuity tests for this age group. The authors point out correctly that the use of a 'crowded' acuity test is important, and cite, as alternatives to the Cambridge Crowding Cards (used in this study) the LEA and HOTV tests. Both the HOTV and LEA tests measure the extent of crowding by using a linear array. This can present a problem to young children of keeping their place on a line, and will result in reduced crowding for the end letter (or end picture) on each line of the chart. These problems are overcome in the Cambridge Crowding Cards in which each test card shows a single central letter, surrounded by crowding letters on all sides, with the child only having to understand what letter is in the 'middle' or 'centre' of the array and point to that letter on the matching board. Perhaps some comment on these advantages may be an appropriate addition to the paper.

- We have clarified that the reference to crowded acuity tests was indeed to the "single" crowded tests, in which a single letter is surrounded by 4 bars and each letter is presented individually to the child, a configuration that eliminates the problem of letters/symbols at the "end of the line" having less crowding. (Section 2.3, page 17). "We do not expect the results to differ depending on the type of acuity test used, as long as an age-appropriate *crowded* acuity test is used that uses matching and presents each letter/symbol individually (e.g., HOTV Hand-Held 50% Crowded Book; Lea Symbols Crowded Symbol Book), because single-letter acuity tests without crowding are less sensitive in identifying individuals with amblyopia."

VERSION 2 – REVIEW

REVIEWER	Bruce Moore OD New England College of Optometry Boston, USA
REVIEW RETURNED	23-Aug-2019

GENERAL COMMENTS	The authors did a great job in addressing the relatively minor comments by the reviewers. I appreciate the clarity of their responses. I think this paper is ready to go, and I can not suggest any other comments that would hinder timely publication. This paper makes an important contribution to our understanding of this complex issue. I look forward to its publication and the discussion that will ensue.

REVIEWER	Professor Janette Atkinson University College London, London, UK
REVIEW RETURNED	30-Aug-2019

GENERAL COMMENTS	1. ABSTRACT : I would suggest the following rewording to the Conclusions to clarify the results 'A school-based screening program correctly identified 84% of those kindergarten children who were found to have a visual problem on screening. Additional analyses revealed how accuracy varies with different combinations of screening tools and referral criteria'.

	This makes it clear that you are not stating that there are visual problems in 84% of the total population who have been screened. 2. Page 17, lines 27-31 Please modify this sentence to add on line 31 ‘Cambridge Crowding Cards, so that line 31 reads ‘letter/symbol individually ‘(e.g. Cambridge Crowding Cards, HOTV Hand-Held 50% Crowded Book; Lea Symbols Crowded Symbol Book).

VERSION 2 – AUTHOR RESPONSE

Second response to reviewers for manuscript: “Choosing appropriate tools and referral criteria for vision screening of 4- and 5-year-old children in Canada: a quantitative analysis”.

The following revisions have been made to the final version of the manuscript, as suggested by Reviewer #2.

Reviewer 2:

1. ABSTRACT : I would suggest the following rewording to the Conclusions to clarify the results

‘A school-based screening program correctly identified 84% of those kindergarten children who were found to have a visual problem on screening. Additional analyses revealed how accuracy varies with different combinations of screening tools and referral criteria’.

This makes it clear that you are not stating that there are visual problems in 84% of the total population who have been screened.

2. Page 17, lines 27-31 Please modify this sentence to add on line 31 ‘Cambridge Crowding Cards, so that line 31 reads

‘letter/symbol individually ‘(e.g. Cambridge Crowding Cards, HOTV Hand-Held 50% Crowded Book; Lea Symbols Crowded Symbol Book).