PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (see an example) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below. Some articles will have been accepted based in part or entirely on reviews undertaken for other BMJ Group journals. These will be reproduced where possible.

ARTICLE DETAILS

| TITLE (PROVISIONAL) | Exploratory Application of the Ages and Stages (ASQ™) Child Development Screening Test in a Low Income Peruvian Shantytown Population |
|---------------------|-----------------------------------------------------------------------------------------------------------------------------|
| AUTHORS             | Oberhelman, Richard; Kyerematen, Victoria; Hamb, Averine; Cabrera, Lilia; Bernabe-Ortiz, Antonio; Berry, Susan               |

VERSION 1 - REVIEW

| REVIEWER          | Scharf, Rebecca, University of Virginia                                      |
|-------------------|----------------------------------------------------------------------------|
| REVIEW RETURNED   | 31-Oct-2013                                                                |

THE STUDY

I appreciate the work of the authors. I think looking for developmental screening tests and assessments that can be applied across languages and cultures around the world is tremendously important. In order to promote early childhood development, we need to be able to measure it and to identify those children most in need of intervention. This is an important topic and worthy of study.

There are several concerns that could be considered:
1. I think the N for the study is low. Administering the Ages and Stages Questionnaire is fairly quick, and I think recruiting more subjects would be worthwhile.
2. I was taught in Public Health School that something is either significant or not. I don't think I would consider p-values over 0.05 significant, and the paper is based on non-significant values.
3. I think the conclusion that it was successfully applied may be broad. It was given, but not compared with any other test, and it didn't point out any significance differences to show which children should be further assessed.
4. Perhaps the authors can adapt the test first (i.e. take out the irrelevant snowman and change to a picture of something culturally relevant) and then pilot and administer the assessment. Since they note that it doesn't appear to be valid in this context, it may be worth adapting, and then completing the study.

Specific concerns:
Introduction:
Page 5, Line 12- the verb should be have, not has (studies have) Could consider re-wording the sentence there are many clauses and it is a bit unclear
Page 5, line 25: screening tests don't diagnose, they screen, so I don't think we can say “prior to screening tests, diagnoses were reliant...”
I think paragraph 3 in the introduction begins with a run-on sentence. The discussion of the Bayley and Denver seems a bit haphazard—perhaps could explain rationale for mentioning them more clearly.
Page 6 line 10: I think there should be a “the” before the 1980's...
Page 6, line 24: The paragraph that begins "this study is an exploratory examination" begins by talking about the ASQ, then gives a little background, but then wanders into the study in Brazil that didn't use the ASQ. I think the thoughts are tangential and could be written more clearly. The paragraphs could have an opening sentence and main idea, and be coherent in thought. Could complete a more thorough literature review.

Methods:
Page 7, line 3: the study makes a vague reference to an "ongoing surveillance project" from which this data was collected, but doesn't mention the name of the study, how it is funded, or give any details. It might be worth giving more details about the parent study, population, other assessments, etc.
Page 7, line 14: why use CDC growth curves instead of WHO? Why was 90% chosen? Need more references.
Line 16: Should "cutoff" be cut off or cut-off?

Results:
The only significant result is that older children were more likely to have a suspect result on the problem solving scale.
I think I would present uniform results in Table 2. Presenting only the data that had the lowest p-values (even though not significant) may not be presenting a full picture? I'm not sure why only 3 of 5 subscales were presented. Under each heading (total ASQ and each subscale), I would consider including the same categories. (Why age for some, gender for some, weight for some? it is not consistent).

Discussion:
I would be cautious discussing non-significant results as if they are significant.
Page 9, line 32: may want to avoid "impacts" and "affect" when discussing associations.
It may be helpful to have an outcome in mind for the study, and then test to see if the ASQ predicted that outcome.

REVIEWER
Jane Squires
University of Oregon
Eugene, Oregon
USA
REVIEW RETURNED 07-Nov-2013

GENERAL COMMENTS
Selecting outcome measures for public health and other studies in low income countries is often a challenge. Using measures developed in the US and elsewhere and studying their psychometric properties in the new setting is one alternative that is often used. This manuscript described a pilot exploratory study in which the ASQ questionnaires (including the ASQ-2 and the ASQ-3) were given to 129 children from a low-income community in Peru and demographic data were collected for analysis on the correlations between three risk factors and the outcomes of screening.

In general the manuscript is well written and describes a straightforward preliminary study using a developmental screening questionnaire to measure child outcomes related to risk factors such as nutritional status. I would suggest the additional of a couple of
Regarding methods and measures, I would like to read some additional detail about recruitment. I'm assuming this is a convenience sample. Were any children excluded from the study? How did recruitment take place? Did testing take place in a clinic setting? Did the Peruvian psychologist confer with the mother at all if he/she had questions on the ASQ that couldn't be answered from the sample of behavior observed in the clinic? Was any attempt made at parent input in the completion process?

Regarding results, I suggest clarifying that "suspect" results mean that child's scores were below the cutoff score established on the ASQ for the US population (if that's correct). Were 3rd edition cut scores used for 3rd edition ASQ and 2nd edition cut scores for 2nd edition? If so, I'd advise using 3rd edition cut scores for both.

In the discussion, I think it important to mention that the ASQ authors found few significant differences between the 2 and 3rd ASQ editions, supporting the logic to combine the two samples (only minor wording revisions were made and there was a 1 item substitution on the 12 month communication domain).

The limitations of the study do need to be emphasized--that this was a convenience sample and that the sample was small across intervals. In addition, there needs to be further validation in Peru to study the cut off scores and necessary adaptations that need to be made. At a minimum, a large normative sample needs to be collected to compare the Peruvian cut scores with US as well as other countries. Ideally, validity and reliability studies will also be conducted.

In general, I think this is a preliminary feasibility study that will have interest to those conducting research with young children in resource poor countries. Especially in the public health arena, there are often few options for assessments and no funding to devote to studying psychometrics of outcome measures. Sharing results like these may encourage others to conduct and report on similar studies.
VERSION 1 – AUTHOR RESPONSE

1. Reviewer Name Scharf, Rebecca
Institution and Country: University of Virginia, USA
Please state any competing interests or state 'None declared': None

I appreciate the work of the authors. I think looking for developmental screening tests and assessments that can be applied across languages and cultures around the world is tremendously important. In order to promote early childhood development, we need to be able to measure it and to identify those children most in need of intervention. This is an important topic and worthy of study.

There are several concerns that could be considered:
1. I think the N for the study is low. Administering the Ages and Stages Questionnaire is fairly quick, and I think recruiting more subjects would be worthwhile.
2. I was taught in Public Health School that something is either significant or not. I don't think I would consider p-values over 0.05 significant, and the paper is based on non-significant values.
3. I think the conclusion that it was successfully applied may be broad. It was given, but not compared with any other test, and it didn't point out any significance differences to show which children should be further assessed.
4. Perhaps the authors can adapt the test first (i.e. take out the irrelevant snowman and change to a picture of something culturally relevant) and then pilot and administer the assessment. Since they note that it doesn't appear to be valid in this context, it may be worth adapting, and then completing the study.

RESPONSE: The study N is over 100 children, which required considerable time and effort to recruit, and our available time commitments for faculty and student investigators do not allow us to extend this study and include additional participants. While many p values reported are slightly greater than 0.05, the authors believe that these may represent biologically relevant differences even if they do not meet the statistical cut off value of 0.05, which is of course based on a statistical convention. If additional support becomes available in the future to investigate these findings with a larger sample, the precedent of these data, even with sample size limitations, will be useful for directing these future research efforts. For additional clarity, values in the Table that are not significant at the p<0.05 level are labeled as NS. Comments in #3 above have been taken into consideration in a revised discussion, and comments in #4 can be considered in subsequent research efforts but are impossible to incorporate in this study.

Specific concerns:
Introduction:
Page 5, Line 12- the verb should be have, not has (studies have) Could consider re-wording the sentence there are many clauses and it is a bit unclear
Page 5, line 25: screening tests don't diagnose, they screen, so I don't think we can say "prior to screening tests, diagnoses were reliant...."
I think paragraph 3 in the introduction begins with a run-on sentence.
The discussion of the Bayley and Denver seems a bit haphazard- perhaps could explain rationale for mentioning them more clearly.
Page 6 line 10: I think there should be a "the" before the 1980's...
Page 6, line 24: The paragraph that begins "this study is an exploratory examination" begins by talking about the ASQ, then gives a little background, but then wanders into the study in Brazil that didn't use the ASQ. I think the thoughts are tangential and could be written more clearly. The paragraphs could have an opening sentence and main idea, and be coherent in thought. Could complete a more thorough literature review.

RESPONSE: These editorial comments have all been considered and changes based on them are incorporated into the revised manuscript. The additional references recommended by reviewer 2 have been incorporated into an expanded literature review, as suggested.

Methods:
Page 7, line 3: the study makes a vague reference to an "ongoing surveillance project" from which this data was collected, but doesn't mention the name of the study, how it is funded, or give any details. It might be worth giving more details about the parent study, population, other assessments,
etc.
Page 7, line 14: why use CDC growth curves instead of WHO? Why was 90% chosen? Need more references.
Line 16: Should “cutoff” be cut off or cut-off?
RESPONSE: Additional information on the surveillance study used to identify potential study participants has been added. CDC growth standards have been widely used for nutritional surveillance in our study population in the past, facilitating data analyses, and while the WHO standards are often advocated because the population includes more exclusively breast fed infants and mothers who did not smoke, the highly selected nature of that reference population is unlikely to be more representative of the Peruvian study population than the CDC reference standards. 90% of the mean weight for age value is the cut off for first degree malnutrition using the Gomez criteria for malnutrition widely applied in Latin America, which is why that value was used for identifying undernourished children (reference included in revised manuscript).

Results:
The only significant result is that older children were more likely to have a suspect result on the problem solving scale.
I think I would present uniform results in Table 2. Presenting only the data that had the lowest p-values (even though not significant) may not be presenting a full picture? I'm not sure why only 3 of 5 subscales were presented. Under each heading (total ASQ and each subscale), I would consider including the same categories. (Why age for some, gender for some, weight for some? it is not consistent).
RESPONSE: The authors do not believe that greatly expanding Table 2 to show the many results that were clearly not different between groups, mostly with p values ranging from 0.3 to 1.0, would merit the additional space needed to show that data. As described in the text, only variables with trends toward significance at the p<0.05 level (i.e. those with p values <1.0) are shown. The revised presentation in the table and discussion makes it clear that many of these do not meet the p<0.05 significance threshold, but yet may merit note because they could reflect biological differences that are worth exploring further with additional resources in the future.

Discussion:
I would be cautious discussing non-significant results as if they are significant.
Page 9, line 32: may want to avoid “impacts” and “affect” when discussing associations.
It may be helpful to have an outcome in mind for the study, and then test to see if the ASQ predicted that outcome.
RESPONSE: See notes above. Wording choices are also revised as suggested.

2. Reviewer Name: Jane Squires
Institution and Country University of Oregon
Eugene, Oregon USA
Please state any competing interests or state ‘None declared’: I am an author of the ASQ and receive some royalties for its publication.

Selecting outcome measures for public health and other studies in low income countries is often a challenge. Using measures developed in the US and elsewhere and studying their psychometric properties in the new setting is one alternative that is often used. This manuscript described a pilot exploratory study in which the ASQ questionnaires (including the ASQ-2 and the ASQ-3) were given to 129 children from a low-income community in Peru and demographic data were collected for analysis on the correlations between three risk factors and the outcomes of screening.

In general the manuscript is well written and describes a straightforward preliminary study using a developmental screening questionnaire to measure child outcomes related to risk factors such as nutritional status. I would suggest the additional of a couple of references in the introduction pertinent to the study; Hix-Small, H., Marks, K., Squires, J., & Nickel, R. (2007). Impact of implementing developmental screening at 12 and 24 months in a pediatric practice. Pediatrics, 120(2), 381–389;
and Filgueiras, A., Pires, P., Maissonette, S., & Landeira-Fernandez, J. (2013). Psychometric properties of the Brazilian-adapted version of the Ages and Stages Questionnaire in public child daycare centers. Early Human Development, 89, 561–576.

RESPONSE: These references have been added as suggested.

Regarding methods and measures, I would like to read some additional detail about recruitment. I'm assuming this is a convenience sample. Were any children excluded from the study? How did recruitment take place? Did testing take place in a clinic setting? Did the Peruvian psychologist confer with the mother at all if he/she had questions on the ASQ that couldn't be answered from the sample of behavior observed in the clinic? Was any attempt made at parent input in the completion process?

RESPONSE: Additional information based on these questions has been added to the methods section.

Regarding results, I suggest clarifying that "suspect" results mean that child's scores were below the cutoff score established on the ASQ for the US population (if that's correct). Were 3rd edition cut scores used for 3rd edition ASQ and 2nd edition cut scores for 2nd edition? If so, I'd advise using 3rd edition cut scores for both.

RESPONSE: Clarifications to these points have been added in the revised manuscript.

In the discussion, I think it important to mention that the ASQ authors found few significant differences between the 2 and 3rd ASQ editions, supporting the logic to combine the two samples (only minor wording revisions were made and there was a 1 item substitution on the 12 month communication domain).

RESPONSE: This comment regarding the logic of combining the results of the ASQ2 and 3 has been added.

The limitations of the study do need to be emphasized--that this was a convenience sample and that the sample was small across intervals. In addition, there needs to be further validation in Peru to study the cut off scores and necessary adaptations that need to be made. At a minimum, a large normative sample needs to be collected to compare the Peruvian cut scores with US as well as other countries. Ideally, validity and reliability studies will also be conducted.

RESPONSE: These comments have been addressed in the revised manuscript with additional comments about the limitations of the study and the need for larger studies to assess potential differences in an expanded, normative sample.

In general, I think this is a preliminary feasibility study that will have interest to those conducting research with young children in resource poor countries. Especially in the public health arena, there are often few options for assessments and no funding to devote to studying psychometrics of outcome measures. Sharing results like these may encourage others to conduct and report on similar studies.

RESPONSE: The authors acknowledge and agree with this comment.
GENERAL COMMENTS

I suggest a few minor edits to clarify meaning. First, in paragraph before results, middle of paragraph: "For each scale, results are reported as normal or suspect; SUSPECT INDICATES THE CHILD’S SCORE WAS BELOW THE CUT OFF SCORE ESTABLISHED FOR THE ASQ-3 IN THE US POPULATION. SOME MILESTONE ATTAINMENT....

Second, in first paragraph of results section, According to the wealth index, the population was fairly equally divided BETWEEN 2 GROUPS, LOWEST AND HIGHEST WEALTH ATTAINMENT.

Third, in the discussion section, regarding the snowman question. It is indeed a contribution to the number of children identified. However children would have to have missed several other items in the domain to have a score indicating potential developmental problems. Therefore this item may have contributed to increased identification rate but was ot the sole cause. Perhaps a sentence could be added that clarifies that this item was not the only reason for children not passing this domain.

RESPONSE: The suggested wording was incorporated.

Second, in first paragraph of results section, According to the wealth index, the population was fairly equally divided BETWEEN 2 GROUPS, LOWEST AND HIGHEST WEALTH ATTAINMENT.

RESPONSE: The suggested wording was incorporated.

Third, in the discussion section, regarding the snowman question. It is indeed a contribution to the number of children identified. However children would have to have missed several other items in the domain to have a score indicating potential developmental problems. Therefore this item may have contributed to increased identification rate but was ot the sole cause. Perhaps a sentence could be added that clarifies that this item was not the only reason for children not passing this domain.

RESPONSE: A clarifying sentence was added as suggested, reading “While the ‘snowman question’ may have contributed to the high rate of suspect results on the problem solving scale, this question was not the sole reason for the results observed since children would have to miss several other items in that domain to fall into the “suspect” category.”