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)	Feasibility and validity of the Health Status Classification System-Preschool (HSCS-PS) in a large community sample: The Generation R Study
AUTHORS	Fang, Xinye; BAI, GUANNAN; Windhorst, Dafna; Feeny, David; Saigal, S; Duijts, Liesbeth; Jaddoe, Vincent; Hu, Shanlian; Jin, Chunlin; Raat, Hein

VERSION 1 – REVIEW

REVIEWER	Mark DeBoer University of Virginia, USA
REVIEW RETURNED	21-Mar-2018

GENERAL COMMENTS	Dr. Fang et al present data from an analysis of a scale for assessing health-related quality of life. Such patient-reported outcomes have been receiving more attention, and use of such a form in preschool children would be helpful. I do have concerns. The usual approach to validation is to compare a new tool (or in this case, a tool being used in a new setting) to some accepted "gold standard" and demonstrate that the new tool corresponds acceptably to the gold standard. In the current study, the authors make multiple hypotheses that some parts of their score or the whole score will correlate with other items: a. Particular domains of the score with other parts of the score. b. The sum of the score with specific medical diagnoses (which presumably they hypothesize will have lower function/quality of life). In each case, when the authors demonstrate that their hypothesis about a particular correlation was true, they take this association as validation of the score's performance. But in each case, the result says more about their own hypothesis than about a validation that these two parts of the score are truly assessing similar concepts in a valid manner. In the case of comparison of domains with specific other questions, this could be merely showing that pessimistic parents who rate their child as >1 in a certain domain are likely to rate that child as >1 in another domain. Indeed, many of the associations that they hypothesize to be lower were nevertheless significantly associated (Table 2). In the case of the low birthweight and preterm children the authors should cite literature that corroborates their assumption that QOL is lower in these groups.

	Overall, I do not think that it is valid to only compare a score to itself and to how one thinks certain diagnoses should be. The authors need to cite external evidence, if they have not used a gold standard in their own analysis and should discuss these limitations further.

REVIEWER	Luis Rajmil Retired. Spain
REVIEW RETURNED	04-Apr-2018

GENERAL COMMENTS	The study presents the feasibility, concurrent validity and discriminant validity of the Health Status Classification System-Preschool (HSCS-PS) in a large community sample of 3 years old children in the Netherlands. The study is really well presented and the conclusions are plausible and understandable. In my opinion the paper needs some minor revisions: 1) Authors have acknowledged on the limitations of the instrument. One of these limitations is related to the generalizability of the results. Non-participants were more often children from vulnerable families, such as single parents, parents with a low educational level, and non-Dutch parents. These groups of children usually report more health problems and worst health status than their counterparts from the wealthier groups. Besides this limitation is usual in these types of studies: could it have had any impact on the results, for example in the high ceiling effect? Although the instrument has been developed from the econometric point of view: has it been proposed to carry out focus groups to review the content of the instrument with specific population subgroups in order to check the understanding, and to try to reduce the number and contents of the items? The authors could briefly discuss these questions in the limitations of the study. 2) Would it be possible to present 95%CI of Spearman cc for the concurrent validity? Cc are very frequent statistically significant when analyzing large samples like in the present study. 3) Related to the previous question: do authors estimated the statistical power to detect differences or the necessary sample size regarding hypotheses for concurrent and discriminant validity? It would be good to add a sentence regarding this issue in the discussion section

REVIEWER	Lai, P.C. The University of Hong Kong, Hong Kong, China
REVIEW RETURNED	23-Aug-2018

GENERAL COMMENTS	This is a study about HSCS-PS from the Generation R data. The authors have included essential details in the supplementary materials. While the study has clearly defined objectives, its coverage is somewhat limited in scope. The study design involved many variables and simplistic measures. The outcomes are discussed in straightforward but rather bland manner. In other words, the presentation is objective, factual, and dry. The methods section could be better integrated to improve structural coherence. The study limitations are buried in the body of the discussion section. The last two paragraphs could serve as the conclusion of the study.

VERSION 1 – AUTHOR RESPONSE

Reviewer's Comments

Reviewer 1

1. Dr. Fang et al present data from an analysis of a scale for assessing health-related quality of life. Such patient-reported outcomes have been receiving more attention, and use of such a form in preschool children would be helpful. I do have concerns.

The usual approach to validation is to compare a new tool (or in this case, a tool being used in a new setting) to some accepted "gold standard" and demonstrate that the new tool corresponds acceptably to the gold standard. In the current study, the authors make multiple hypotheses that some parts of their score or the whole score will correlate with other items:

a. Particular domains of the score with other parts of the score.

b. The sum of the score with specific medical diagnoses (which presumably they hypothesize will have lower function/quality of life).

In each case, when the authors demonstrate that their hypothesis about a particular correlation was true, they take this association as validation of the score's performance. But in each case, the result says more about their own hypothesis than about a validation that these two parts of the score are truly assessing similar concepts in a valid manner. In the case of comparison of domains with specific other questions, this could be merely showing that pessimistic parents who rate their child as >1 in a certain domain are likely to rate that child as >1 in another domain. Indeed, many of the associations that they hypothesize to be lower were nevertheless significantly associated (Table 2).

Author's response: Dear Reviewer, thank you for your comments. In this study there was no separate measurement of HRQOL that can serve as 'gold standard' in order to evaluate the concurrent validity. In this study however, two single-items were included regarding the measurement of 'General health', and 'Behavior'. These items are separate and additional to the original 10 HSCS-PS domains. According to Macias et al., a single-item self-rating 'General health' can serve as valid, reliable and sensitive measure for research purposes, although more research is recommended. According to

Ahmad et al., a single-item self-rating ‘Mental health’ can serve as an appropriate population mental health measure. Ahmad et al. recommend more research to examine relationships with future mental health. Therefore, in the absence of a ‘gold standard’, we propose using the two additional (separate) single-items for ‘General health’ and ‘Behavior’ only as a first careful assessment of the concurrent validity of the CHSCS-PS. In the revised manuscript, we address this issue in the Introduction section, the Methods section, and the Discussion section; also the Abstracts and Article Summary are adapted.

We agree with the reviewer that the ‘response tendency’ of a (for example) optimistic or a pessimistic parent may influence the answers to all items in the questionnaire towards one direction. This might contribute to the association between hypothesized parallel constructs in the study. We add this discussion as a limitation to the Discussion section of the revised manuscript. We agree with the reviewer that the evaluation of the Discriminant validity is also based on the a-priori hypothesis that certain subgroups with a ‘medical’ condition are expected to have lower HRQOL scores. Therefore, as the reviewer suggests, in the revised manuscript (Discussion section), we will discuss the limitations of this study, and we will recommend a future study including a second accepted ‘gold standard’ measure of HRQOL to further evaluate the validity.

In the revised manuscript we added the following sentences:

Abstract. Line 72-74: *“ In the absence of another HRQOL measure, this study uses two additional single-item measured ‘General health’ and ‘Behavior’ as a first step to evaluate concurrent validity of the HSCS-PS.”*

Abstract. Line 83-84: *“Concurrent validity: HSCS-PS domains correlated better with hypothesized parallel additional domains than with other non-hypothesized original domains.”*

Article Summary. Line 101-103: “*In the absence of an accepted ‘gold standard’ in this study, we recommend to assess the association between HSCS-PS with one or more other established measures of HRQOL in a future study.*”

Introduction. Line 140-142.: *“In the absence of another HRQOL measure, this study uses two single-item questions regarding ‘General health’[17] and ‘Behavior’ [18] as a first step to evaluate concurrent validity of the HSCS-PS.”*

Introduction. Line 148-150*: "and (3) as a first step regarding the concurrent validity by evaluation of the correlations between the original HSCS-PS scores and the 'General health' and 'Behavior' single-item measures."*

Methods. Line 175-178: *"In addition, Saigal et al. proposed two additional parent-reported single-item questions regarding 'General Health' and 'Behavior', given the relatively high prevalence of general health and behavior problems among the very-low-birth-weight (VLBW) infants.[30, 31]"*

Methods. Line 247-260:

"Concurrent Validity

In the absence of a 'gold standard' measure of HRQOL, as a first step to evaluate the concurrent validity of the 10-domains HSCS-PS, it was assessed whether specific HSCS-PS domains correlated better with their assumed 'parallel' additional single-item measures of 'General health' and/or, 'Behavior' than with a 'non-parallel' measure. Considering the non-normal distribution of the data, Spearman rank correlation was applied. We calculated bootstrapped 95% confidence intervals for Spearman correlation coefficients. When (a) 95% confidence interval is not 'across 0'; and (b) the p value < 0.05, the correlation coefficient was regarded as statically significant. We hypothesized relatively high correlation coefficients between the following 'parallel' pairs of a HSCS-PS-domain/single-item parent-rated measure (in italics): 'Pain and discomfort'/'General health'; 'Self-care'/'Behavior'; 'Emotion'/'Behavior'; 'Learning and remembering'/'Behavior'; 'Thinking and problem solving'/'Behavior'; and we hypothesized the correlation coefficients for all other pairs were hypothesized to be lower."

Discussion. Line 365-371: *"It should be noted that relatively little is known about the acceptance and validity of parent-report single-items to describe 'General health' and 'Behavior/Mental health' of children compared to the body of knowledge regarding the validity of such measures in adult populations. [26, 27] Therefore, in the future, we recommend the concurrent validity of the HSCS-PS should be evaluated by comparing it with an accepted 'gold standard' HRQOL measure such as the Infant and Toddler Quality of Life Questionnaire (ITQOL).[6] The evaluation of the concurrent validity of the 10-domains HSCS-PS in this study is a first step and results should be interpreted with caution. "*

In the case of the low birth weight and preterm children the authors should cite literature that corroborates their assumption that QOL is lower in these groups.

Overall, I do not think that it is valid to only compare a score to itself and to how one thinks certain diagnoses should be. The authors need to cite external evidence, if they have not used a gold standard in their own analysis and should discuss these limitations further.

Author's Response: We added external literature in the revised Introduction section showing that proxy-reported HRQOL of preschool children was lower when children had the health condition which was studied in the present study.

Introduction. Line 135-139: *"Previous studies have shown that children with the above-mentioned health condition were reported by their parents or caregivers with relatively low HRQOL.[17-25] For example, the parent-reported HRQOL of preschool children born preterm or born with a very low birth weight was lower than HRQOL of those who were born not born in preterm or with a low birth weight.[17-20]"*

Regarding the gold standard, we fully agree with the reviewer and have addressed this issue in the revised Introduction, Methods, Discussion sections. Please see the above sentences in the first response.

Reviewer 2

The study presents the feasibility, concurrent validity and discriminant validity of the Health Status Classification System-Preschool (HSCS-PS) in a large community sample of 3 years old children in the Netherlands. The study is really well presented and the conclusions are plausible and understandable. In my opinion the paper needs some minor revisions:

1. Authors have acknowledged on the limitations of the instrument. One of these limitations is related to the generalizability of the results. Non-participants were more often children from vulnerable families, such as single parents, parents with a low educational level, and non-Dutch parents. These groups of children usually report more health problems and worst health status than their counterparts from the wealthier groups. Besides this limitation is usual in these types of studies: could it have had any impact on the results, for example in the high ceiling effect?

Author's response: Dear Reviewer, thank you for your comments. We fully agree with you that the non-participants were more often children from vulnerable families which may impose an impact on results. We have addressed this issue in the revised Discussion section. Please see the texts from Line 395-400: *"Third, in our study, the non-participants were children from vulnerable families, who more often had single parent, and whose parents more often had lower educational level or had an immigrant background. These children may have more health conditions/problems than their counterparts from non-vulnerable families. This issue may impose an impact on results. For instance, the high ceiling effect may be caused by the relatively better health status of the participants. In addition, the generalizability of results in the present study may be limited due to this issue."*

2. Although the instrument has been developed from the econometric point of view: has it been proposed to carry out focus groups to review the content of the instrument with specific population subgroups in order to check the understanding, and to try to reduce the number and contents of the items? The authors could briefly discuss these questions in the limitations of the study.

Author's response: According to the design paper Saigal et al.(published in 2005), no focus group discussion was conducted during the development of the instrument for preschool children. The draft HSCS-PS instrument was applied to approximately 80 children in Canada in order to check the content validity. In our revised manuscript we added sentences in the revised Methods section to describe the development of HSCS-PS in more details, also we addressed the issue of no focus group discussion in the revised Discussion section. Please see the sentences below.

Methods. Line 174-181: "*The HSCS-PS is a parent reported health status questionnaire applicable to 2.5-5 year-olds which consists of 10 mutually exclusive domains, based on the Health Utility Index (HUI).[11] In addition, Saigal et al. proposed two additional parent-reported single-item questions regarding 'General Health' and 'Behavior', given the relatively high prevalence of general health and behavior problems among the very-low-birth-weight (VLBW) infants.[30, 31] The HSCS-PS was initially applied to approximately 80 children across Canada by pediatricians and neonatologists regarding the structured and qualitative feedback. After several rounds of refinements, the final version contains 10 domains each with 3-5 levels, and the two additional items. (see S1 Table)."*

Discussion-Methodological considerations. Line 416-420: *"Fifth, we would like to note that regarding the procedure of developing the HSCS-PS, items were mainly derived from the HUI system and additionally two new items were based on experts' opinion. Qualitative studies, such as using focus group interviews have not been mentioned in this procedure; we recommend that qualitative research may be applied in the future, for example, to reduce the number of items, or to evaluate the content of the items."*

3. Would it be possible to present 95%CI of Spearman cc for the concurrent validity? Cc are very frequent statistically significant when analyzing large samples like in the present study.

Author's response: Dear Reviewer, thank you for your comment. We fully agree with you that in the large samples, correlation coefficients are very often statistically significant, therefore, we have added the information of 95% confidence interval (CI) of Spearman correlation coefficient. Please see the texts from line 252 to 255 in the revised Methods section: *"We calculated bootstrapped 95% confidence intervals for Spearman correlation coefficients. When (a) 95% confidence interval is not 'across 0'; and (b) p value < 0.05, the correlation coefficient was regarded as statistically significant."* Also, please see the revised Table 4 as blow and (also) in the revised manuscript. We want to address that the 95% CIs are cross 0 regarding the correlation between 'Vision' and 'General Health', and correlation between 'Hearing' and 'Behavior'. In these two cases, we think the correlation is not statistically significant, even though the p value is less than 0.01.

Table4. Concurrent validity of the HSCS-PS assessed by Spearman correlations between original HSCS-PS domains and two additional domains (n=4546)*

HSCS-PS domains	General health	Behavior
Vision	0.04(-0.003, 0.094)	0.02 (-0.011, 0.060)
Hearing	0.09** (0.039, 0.143)	0.04 (-0.002, 0.085)
Speech	0.08** (0.047, 0.111)	0.09** (0.059, 0.126)
Mobility	0.13** (0.066, 0.187)	0.06** (0.015, 0.106)
Dexterity	0.11** (0.038, 0.178)	0.07** (0.013, 0.129)
Self-care	0.09** (0.051, 0.129)	*0.17** (0.129, 0.218)*
Emotion	0.00 (-0.016, 0.033)	*0.16*** (0.088, 0.221)
Learning and remembering	0.11** (0.051, 0.178)	*0.16***(0.091, 0.228)
Thinking and problem solving	0.11** (0.056, 0.176)	*0.18*** (0.123, 0.245)
Pain and discomfort	*0.21*** (0.161, 0.250)	0.08** (0.040, 0.109)

Values presented in this table are values of Spearman correlation coefficient (CC) and 95% confidence interval (CI) of Spearman's CC.
*Correlations with predefined related general health/behavior are in italics; other (spurious) are in standard font.

** When (a) 95% confidence interval is not 'across 0'; and (b) p value < 0.05, the correlation coefficient was regarded as statistically significant.
HSCS-PS = Health Status Classification System-Preschool

4. Related to the previous question: do authors estimated the statistical power to detect differences or the necessary sample size regarding hypotheses for concurrent and discriminant validity? It would be good to add a sentence regarding this issue in the discussion section

Author's response: We agree with the reviewer that statistical power to detect differences and significant associations is an important consideration. To our knowledge there are no strict guidelines for the size of such validation studies with multiple comparisons and multiple associations that are studied. Because our study was embedded in a large cohort study (Generation R) with 4546 children/parents in the analyses, which is relatively high for a validation study, we consider that the power was sufficient. Although extreme subgroups in the evaluation of Discriminative ability (birth weight < 1500 grams, n=28; gestational age <32 weeks, n=31) were relatively small, in both comparisons (<1500 grams; < 32 weeks) 6 out of 13 differences were statistically significant. All other comparisons were based on much larger subgroups (n=104 to n=4307). This indicates that the statistical power was sufficient for the purpose of this validation study.

In the revised manuscript we added to the paragraph regarding methodological considerations in the Discussion section, lines 388-394: *"Second, no formal power calculations were made with regard to the validation study, given multiple comparisons and studies of associations. However, the size of the population for analysis (n=4526) is relatively large for a validation study; therefore many associations, even with a small effect size, were statistically significant. The smallest subgroups regarding the evaluation of discriminative validity (birth weight < 1500 grams, n=28; gestational age <32 weeks, n=31) resulted in almost half of the comparisons being statistically significant. All other subgroups regarding the evaluation of discriminative validity ranged from n=104 up to n=4307."*

Reviewer 3

1. This is a study about HSCS-PS from the Generation R data. The authors have included essential details in the supplementary materials. While the study has clearly defined objectives, its coverage is somewhat limited in scope. The study design involved many

variables and simplistic measures. The outcomes are discussed in straightforward but rather bland manner. In other words, the presentation is objective, factual, and dry. The methods section could be better integrated to improve structural coherence.

Author's response: We thank the reviewer for this comment. Given the comments by all three reviewers, we rewrote the manuscript. In particular, we followed your advice to restructure the Methods and Discussion sections in order to make it more coherent.

2. The study limitations are buried in the body of the discussion section.

Author's response: We restructured the limitation and conclusion of this study. In our revised manuscript we address this in Discussion section using the subtitle "Methodological consideration". Please see the texts from line 381-423.

"*Methodological considerations*

First, in this study, measurements were primarily done using parent questionnaires, including accepted validated instruments such as the Child Behavior Checklist parent questionnaire.[36] Only the birth outcomes were obtained from medical files. 'Reporting tendency' by, for example, 'optimistic' or 'pessimistic' parents may have applied to all measures in the questionnaires and may have induced relatively high statistical associations in this study. For future validation studies we recommend to use as many as possible 'objective' external measures to validate the 10-domains HSCS-PS.

Second, no formal power calculations were made with regard to the validation study, given multiple comparisons and studies of associations. However, the size of the population for analysis (n=4526) is relatively large for a validation study; therefore many associations, even with a small effect size, were statistically significant. The smallest subgroups regarding the evaluation of discriminative validity (birth weight < 1500 grams, n=28; gestational age <32 weeks, n=31) resulted in almost half of the comparisons being statistically significant. All other subgroups regarding the evaluation of discriminative validity ranged from n=104 up to n=4307.

Third, in our study, the non-participants were children from vulnerable families, who more often had single parent, and whose parents more often had lower educational level or had an immigrant background. These children may have more health conditions/problems than their counterparts from non-vulnerable families. This issue may impose an impact on results. For instance, the high ceiling

effect may be caused by the relatively better health status of the participants. In addition, the generalizability of results in the present study may be limited due to this issue."

Fourth, while a utility-based scoring algorithm for HSCS-PS has not yet been developed, a total 'disability score' summing up the scores regarding each of the ten original domains was applied in this study.[16] Two previous studies supported the feasibility and validity of the HSCS-PS total 'disability score' in absence of a utility-based scoring algorithm, which we recommend to be developed in future studies.[15, 16, 47] Given the relative paucity of experience with the HSCS-PS system, no specific guidelines for clinically important differences are available; we recommend such guidelines to be developed. Regarding the Health Utilities Index for patients aged four years and above, it was proposed that a difference of one level within any domain may be interpreted as a clinically important difference.[12] In our case, for example, the subgroup with CBCL 'behavior problems present' and the subgroup with '≥3 chronic/medical conditions' have both a mean total 'disability score' that is more than 1 point (1 level) higher compared to the reference group, which may be interpreted as a clinically important difference. From a statistical point of view, we propose to apply Cohen's effect size (d), and to interpret 0.50 (half a standard deviation) as a meaningful difference. Effect sizes were relatively small in this study, which reflects that the general population in a society with modern and accessible health care is relatively healthy.[38, 48]

Fifth, we would like to note that regarding the procedure of developing the HSCS-PS, items were mainly derived from the HUI system and additionally two new items were based on experts' opinion. Qualitative studies, such as using focus group interviews have not been mentioned in this procedure; we recommend that qualitative research may be applied in the future, for example, to reduce the number of items, or to evaluate the content of the items.

Finally, in the present study, indicators of the reliability of the HSCS-PS, such as test-retest reliability were not evaluated. We recommend assessing this in future studies in the large varied community population."

3. The last two paragraphs could serve as the conclusion of the study.

Author's response: We have combined the sentences regarding strengths of this study with the last paragraph into Conclusion section. The sentences regarding limitations of this study has been moved to the Methodological considerations.

Please see texts from line 424 to 435.

"Conclusion

This study is the first to apply and to evaluate the HSCS-PS in a large community sample of preschool children. This is a relevant addition to previous studies among very low birth weight children and children with cerebral palsy. For the assessment of the validity, we applied objectively measured conditions (birth weight, gestational age at birth) in addition to validated parent-reported outcome measures (CBCL). This study supports the feasibility and validity of the HSCS-PS among preschool children in community settings. We recommend developing utility-based scoring algorithms for the HSCS-PS, and conducting empirical studies of what changes are meaningful, as well as repeated studies of reliability and validity in large varied populations with objectively measured, external benchmarks. In the meantime, the HSCS-PS may be used by clinicians and researchers as parent-reported health outcome in addition to clinical outcomes for economic evaluations, and may be used to support the development of value-based health care regarding interventions for preschool children."

VERSION 2 – REVIEW

REVIEWER	Mark DeBoer University of Virginia. USA
REVIEW RETURNED	28-Sep-2018

GENERAL COMMENTS	The authors have satisfactorily addressed my concerns.

REVIEWER	Luis Rajmil Retired. Spain
REVIEW RETURNED	02-Oct-2018

GENERAL COMMENTS	Authors have adequately answered my questions and concerns.