

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)	Strengths and Difficulties Questionnaire: internal validity and reliability for New Zealand pre-schoolers
AUTHORS	Kersten, Paula; Vandal, Alain; Elder, Hinemoa; McPherson, Kathryn

VERSION 1 – REVIEW

REVIEWER	Kareena McAloney-Kocaman Glasgow Caledonian University, Scotland, UK
REVIEW RETURNED	07-Feb-2018

GENERAL COMMENTS	An interesting and well performed analysis of the utility of the SDQ in the context of New Zealand pre-school children. The analysis is suitable, appropriately conducted and well reported. The findings are of interest to the wider academic and practitioner communities, and the paper is presented well.
-------------------------	--

REVIEWER	Jeffrey M. DeVries Technische Universität Dortmund Germany
REVIEW RETURNED	12-Feb-2018

GENERAL COMMENTS	<p>Overall, I believe the research methodology is sound, and the article makes key contributions to our knowledge of the SDQ. More details regarding the methodology are necessary, and a deeper discussion of the theoretical background, implications, and this study's relationship to other SDQ studies will help clarify this study's contribution to the body of research. Below are my specific observations and recommendations to improve the article.</p> <p>The strengths of this article include a relatively rare Rasch analysis on a commonly used instrument (the SDQ), which includes a breakdown of the 5 factors and the DIF of specific items by ethnicity and/or gender. This is important because the SDQ is a commonly used early screening tool for child developmental problems, and research has rarely assessed it via such strenuous statistical methods. The study also used a random sample from a complete population of New Zealand children. Its methodology is clear and sound, although some clarifications are necessary.</p> <p>The article's weaknesses are a general inadequate job of discussing theoretical implications and relating the work to previous work with the SDQ. As a result the rationale for the hypotheses is unclear. Further, careful consideration is needed in the article's discussion regarding limitations and relating the study to previous work.</p>
-------------------------	--

	<p>The theoretical background in the introduction is sparse. This results in unclear rationale for the specific hypotheses proposed (pg 5). More detailed explanations of the theory and past results, and a summary of how these formed your hypotheses would strengthen this greatly.</p> <p>page 4, line 23: More recent estimators in SEM for CFA include more effective categorical estimators which are used in work with SDQ (e.g., Goodman et al 2010 use WLSMV). Also, other past studies have used partial-credit Rasch analyses for the SDQ. This could be worth mentioning here.</p> <p>Page 4, lines 39-45: A citation regarding inadequate vs. acceptable cutoffs for alpha is needed.</p> <p>Page 5: While, the discussion clarifies that it is a standard process in NZ for all children to receive the screening. More discussion about this process is necessary in the methods section.</p> <p>Page 5: If this is part of a before school check, who are the “teacher” raters in the teacher version of the test? The children would not be in school yet, so is this done by their preschool teachers? Please clarify.</p> <p>Page 6: instruments. There are separate versions of the SDQ for 2- to 4- year-olds and 4 years and up, with separate scoring guides. Because this study uses 4- and 5-year-olds, it is important to clarify which version is used. Is this an evaluation of the 4- to 17-year-old version or the 2- to 4-year old version? Furthermore, which English version of the SDQ is used in NZ (I assume the Austral version, but this needs to be clarified)?</p> <p>Page 5: Is there any reason why some of the data may have so many missing values?</p> <p>Page 6: Are the percents for both the parent and the smaller parent-teacher dataset the same?</p> <p>Page 7: It is unclear how many of your 1000 subject sample are part of the cross-informant analysis. Does the 1000 random sample come from the parent only population or the parent-teacher matched population?</p> <p>Page 7, a citation for paragraph 3 is needed.</p> <p>Please provide details about programs used to fit Rasch models. The researchers claim this is the first study to examine cross cultural invariance at the item level in the SDQ, this is not true. See for example Ortuño-Sieera et al 2015 for cross-national comparisons (across European countries) which includes discussion about invariance on specific items. Also, Richter et al (2011) discusses invariance among ethnic groups within a European sample.</p> <p>A greater discussion of the relationship between this paper’s findings and those of past invariance analyses is important. Even if they did not use the same Rasch approach, many other studies have examined invariance across ethnic, national, and/or cultural lines for the SDQ. Do they find the same problematic items as this study?</p> <p>The article needs to clearly discuss the sample age group in the</p>
--	---

	<p>discussion and limitations. The sample is only from 4-5 year olds. As mentioned above, there is a separate version for 2- to 4-year olds. Because the sample is on this cut-off, did you use this version or the version for older children? And many researchers have done invariance and other DIF analyses on the SDQ versions across multiple age groups (even within just the older children's version). While other researchers have concluded full (e.g., DeVries et al, 2017) or partial invariance (e.g., Hagsquist, 2007) over age groups, these studies focused on older age groups. Some discussion of the equivalency (or lack of) between your age groups and older tests should be discussed.</p>
--	---

VERSION 1 – AUTHOR RESPONSE

Editorial Requirements:	Changes made
Please complete and include a STROBE check-list, ensuring that all points are included and state the page numbers where each item can be found: the check-list can be downloaded from here: http://www.strobe-statement.org/?id=available-checklists	We have uploaded this.
Please revise the strengths and limitations (after the abstract) to make it clear whether each point is being presented as a strength or as a limitation.	We have done this within the discussion and the strengths and limitations box.
Reviewer: 1 Name: Kareena McAloney-Kocaman	
An interesting and well performed analysis of the utility of the SDQ in the context of New Zealand pre-school children. The analysis is suitable, appropriately conducted and well reported. The findings are of interest to the wider academic and practitioner communities, and the paper is presented well.	Nil required
Reviewer: 2 Name: Jeffrey M. DeVries	
Overall, I believe the research methodology is sound, and the article makes key contributions to our knowledge of the SDQ. More details regarding the methodology are necessary, and a deeper discussion of the theoretical background, implications, and this study's relationship to other SDQ studies will help clarify this study's contribution to the body of research. Below are my specific observations and recommendations to improve the article.	Our responses are made below for the specific suggestions in relation to the reviewer's statement here.
The strengths of this article include a relatively rare Rasch analysis on a commonly used instrument (the SDQ), which includes a breakdown of the 5 factors and the DIF of specific items by ethnicity and/or gender. This is important because the SDQ is a commonly used early screening tool for child developmental problems, and research has rarely assessed it via such strenuous statistical methods. The study also used a random sample from a complete population of New Zealand children. Its methodology is clear and sound, although some clarifications are necessary.	Nil required
The article's weaknesses are a general inadequate job of discussing theoretical implications and relating the work to previous work with the SDQ. As a result the rationale for the hypotheses is unclear.	We have added an introductory paragraph explaining the importance of screening for emotional and behavioural difficulties in pre-schoolers.

	We have added a brief discussion of classical test theory and modern test theory to provide more theoretical justification for our approach and linked this more clearly to the aims and hypotheses at the end of the background section.
Further, careful consideration is needed in the article's discussion regarding limitations and relating the study to previous work.	We have expanded the strengths and limitations section and throughout the background and discussion have also included additional literature to address this.
The theoretical background in the introduction is sparse. This results in unclear rationale for the specific hypotheses proposed (pg 5). More detailed explanations of the theory and past results, and a summary of how these formed your hypotheses would strengthen this greatly.	We have added an introductory paragraph explaining the importance of screening for emotional and behavioural difficulties in pre-schoolers. We have added a brief discussion of classical test theory and modern test theory to provide more theoretical justification for our approach and linked this more clearly to the aims and hypotheses at the end of the background section.
page 4, line 23: More recent estimators in SEM for CFA include more effective categorical estimators which are used in work with SDQ (e.g., Goodman et al 2010 use WLSMV).	We have added references to different studies in pre-schoolers as well as the reference suggested by the reviewer which concerned older children.
Also, other past studies have used partial-credit Rasch analyses for the SDQ. This could be worth mentioning here.	We used the partial-credit model and have added this detail at the end of the methods section.
Page 4, lines 39-45: A citation regarding inadequate vs. acceptable cutoffs for alpha is needed.	We have added this.
Page 5: While, the discussion clarifies that it is a standard process in NZ for all children to receive the screening. More discussion about this process is necessary in the methods section.	We have provided more information in the methods section.
Page 5: If this is part of a before school check, who are the "teacher" raters in the teacher version of the test? The children would not be in school yet, so is this done by their preschool teachers? Please clarify.	We have provided more information in the methods section.
Page 6: instruments. There are separate versions of the SDQ for 2- to 4- year-olds and 4 years and up, with separate scoring guides. Because this study uses 4- and 5-year-olds, it is important to clarify which version is used. Is this an evaluation of the 4- to 17-year-old version or the 2- to 4-year old version? Furthermore, which English version of the SDQ is used in NZ (I assume the Austral version, but this needs to be clarified)?	Previously we stated that 68% were aged 4 but this was a rounding error in the variable age. We have clarified that the B4SC is carried out when the child is 4 and exceptionally when they are 5; also that 99.6% were aged 4 when the SDQ was completed and 0.4% aged 5. We have also clarified that during the B4SC the Australian version for children aged 2 to 4 is used.
Page 5: Is there any reason why some of the data may have so many missing values?	We have expanded the section in the methods on data quality checking and included possible reasons in the discussion section.
Page 6: Are the percents for both the parent and the smaller parent-teacher dataset the same?	We have provided more detail on demographic variables for the different samples used.
Page 7: It is unclear how many of your 1000 subject sample are part of the cross-informant analysis. Does the	We have clarified in the text that the 1,000 were taken randomly from the full

1000 random sample come from the parent only population or the parent-teacher matched population?	parent population.
Page 7, a citation for paragraph 3 is needed.	We have added a reference
Please provide details about programs used to fit Rasch models.	This information was provided in the 2 nd paragraph of the data analysis section.
The researchers claim this is the first study to examine cross cultural invariance at the item level in the SDQ, this is not true. See for example Ortuño-Sieera et al 2015 for cross-national comparisons (across European countries) which includes discussion about invariance on specific items. Also, Richter et al (2011) discusses invariance among ethnic groups within a European sample. A greater discussion of the relationship between this paper's findings and those of past invariance analyses is important. Even if they did not use the same Rasch approach, many other studies have examined invariance across ethnic, national, and/or cultural lines for the SDQ. Do they find the same problematic items as this study?	We have expanded our discussion as suggested.
The article needs to clearly discuss the sample age group in the discussion and limitations. The sample is only from 4-5 year olds. As mentioned above, there is a separate version for 2- to 4-year olds. Because the sample is on this cut-off, did you use this version or the version for older children?	Previously we stated that 68% were aged 4 but this was a rounding error in the variable <i>age</i> . We have clarified that the B4SC is carried out when the child is 4 and <i>exceptionally</i> when they are 5; also that 99.6% were aged 4 when the SDQ was completed and 0.4% aged 5. We have also clarified that during the B4SC the Australian version for children aged 2 to 4 is used.
And many researchers have done invariance and other DIF analyses on the SDQ versions across multiple age groups (even within just the older children's version). While other researchers have concluded full (e.g., DeVries et al, 2017) or partial invariance (e.g., Hagsquist, 2007) over age groups, these studies focused on older age groups. Some discussion of the equivalency (or lack of) between your age groups and older tests should be discussed.	We have included additional literature as suggested by the reviewer, both when discussing measurement invariance and then linking this to caseness in the discussion.

VERSION 2 – REVIEW

REVIEWER	Jeffrey M. DeVries Technische Universität Dortmund Germany
REVIEW RETURNED	14-Mar-2018
GENERAL COMMENTS	The authors have adequately addressed my concerns from the previous review. I think this article will be a welcome addition to the body of research.