BMJ Open

Strengths and Difficulties Questionnaire: internal validity and reliability for New Zealand pre-schoolers

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-021551
Article Type:	Research
Date Submitted by the Author:	08-Jan-2018
Complete List of Authors:	Kersten, Paula; University of Brighton, School of Health Sciences Vandal, Alain; AUT University, Department of Biostatistics and Epidemiology; Counties Manukau District Health Board, Ko Awatea Health Intelligence and Informatics Elder, Hinemoa; Te Whare Wānanga o Awanuiārangi, School of Graduate Studies McPherson, Kathryn; Health Research Council of New Zealand; AUT University, Centre for Person Centred Research, School of Clinical Sciences
Keywords:	Strengths and Difficulties Questionnaire, validity, reliability, Rasch, pre-school

Strengths and Difficulties Questionnaire: internal validity and reliability for New Zealand pre-schoolers

Paula Kersten[1], Alain C Vandal[2], Hinemoa Elder[3], Kathryn M McPherson[4 5]

[1] School of Health Sciences, University of Brighton, UK

[2] Department of Biostatistics and Epidemiology, AUT University, New Zealand & Ko Awatea Health Intelligence and Informatics, Counties Manukau District Health Board, New Zealand

[3] Te Whare Wānanga o Awanuiārangi, Auckland, New Zealand.

[4] The Health Research Council of New Zealand

[5] Centre for Person Centred Research, School of Clinical Sciences, AUT University, New Zealand.

Corresponding author

Professor Paula Kersten, School of Health Sciences, University of Brighton, Westlain House, Falmer, Brighton BN1 9PH. p.kersten@brighton.ac.uk. Tel +44 1273 643483. Fax: +44 1273 644010.

Abstract

Objectives: This paper examines the internal construct validity, internal consistency and cross-informant reliability of the Strengths and Difficulties Questionnaire (SDQ) in a New Zealand pre-school population across four ethnicity strata (New Zealand European, Māori, Pasifika, Asian).

Design: Rasch analysis was employed to examine internal validity on a subsample of 1,000 children. Internal consistency (n=29,075) and cross-informant reliability (n=17,006) was examined using correlations, intraclass correlation coefficients and Cronbach's Alpha on the full sample available for such analyses.

Setting & participants: Data was utilised from a national SDQ database provided by the funder, pertaining to New Zealand domiciled children aged 4 and 5, and scored by their parents and teachers.

Results: The five subscales do not fit the Rasch model (as indicated by the overall fit statistics); contain items that are biased (differential item functioning) by key variables, suffer from a floor and ceiling effect and have unacceptable internal consistency. After dealing with differential item functioning the Total Difficulty scale does fit the Rasch model and has good internal consistency. Parent/teacher inter-rater reliability was unacceptably low for all subscales.

Conclusion: The five SDQ subscales are not valid and not suitable for use in their own right in New Zealand. We have provided a conversion table for the Total Difficulty scale, which takes account of bias by ethnic group. Clinicians should use this conversion table in order to reconcile differential item functioning by culture in final scores. It is advisable to use both parents and teachers' feedback when considering children's needs for referral of further assessment. Future work should examine if validity is impacted by different language versions used in the same country.

Keywords

Strengths and Difficulties Questionnaire, validity, reliability, Rasch, pre-school

Strengths and limitations of this study

- A key strength of this study is the inclusion of all 4 and 5 year old children in New Zealand for whom an SDQ assessment was available in 2011, resulting in our ability to assess the validity of the tool at the population level and with sufficient power to make sound conclusions.
- We excluded 39% of data as we had some concerns about their quality (it being incomplete or containing multiple inconsistencies).
- We were unable to assess DIF by other key variables that may affect validity, e.g. first language or country of birth, as such data were not available.
- Future work should examine if validity is impacted by different language versions used (in the same country).

Funding

This work was supported by the Ministry of Health of New Zealand (grant number 341088). The funding body has not had input into the design, data collection, analysis, interpretation of data, in the writing of the manuscript, nor in the decision to submit the manuscript for publication.

Acknowledgements

We thank the funder for supporting the study.

Introduction

The Strengths and Difficulties Questionnaire for parents (SDQ-P) and for teachers (SDQ-T) is a tool used worldwide to screen pre-school children's psychosocial attributes (positive and negative behaviours).[1 2 3 4] It consists of 25 items, making up five subscales: Emotional Symptoms, Conduct Problems, Hyperactivity, Peer Problems, and Prosocial Behaviour.[1 2]

The structural validity of the SDQ has been extensively researched using factor analysis (e.g. by [5-7]). A recent systematic review found acceptable to good evidence for the 5-factor SDQ structure, when confirmatory factor analysis (CFA) had been used.[8] A different approach to examining structural validity can be achieved by examining if each of the subscales are unidimensional and fit the Rasch model (i.e. examining internal construct validity).[9] Like CFA, Rasch analysis is a confirmatory approach to examining if items belong to the subscales under investigation. However, there are known limitations of using factor analysis on ordinal scales, including its parametric basis and the emergence of 'difficulty factors', which may spuriously indicate multidimensionality.[10] In addition, factor analysis does not allow detailed investigation of item function regard to targeting, differential item functioning and local dependency between items, whereas Rasch analysis includes such assessments.[11]

Internal consistency of the SDQ-P subscales has been reported in many studies and synthesised in a systematic review.[8] The sample size-weighted average Cronbach's alphas (α) for the five subscales was below the threshold of 0.70 (implying inadequate internal consistency); and for the Difficulty scale α was 0.79 (acceptable for group comparisons but not for individual use).

Inter-rater reliability of SDQ subscales between two parents and between two teachers has previously been found to be acceptable when correlation coefficients were used (between 0.42 and 0.64 for parents and between 0.59 and 0.81 for teachers.[12] Other studies have examined scores between different types of informants (e.g. parent and teacher). The systematic review showed that the sample size-weighted average correlation coefficients generated from these studies were weak to moderate (between 0.25 and 0.45).[8]

The validity and reliability of the SDQ have not previously been examined in New Zealand, which is a country with a sizeable indigenous population (Māori, 15.4%) and immigrant population (25.2% born overseas).[13] New Zealand is a multi-cultural society, impacting upon values, ways of living and languages spoken. It therefore cannot be assumed that measures capturing psychological constructs will have cultural equivalence.[14 15] Indeed, a New Zealand qualitative study has shown that parents from Māori, Pacific Island, Asian, and new immigrant groups questioned the cultural validity of the SDQ.[16] Cultural equivalence therefore needs further investigation.

This study aimed to examine the reliability between parents and teachers (cross-informant reliability), internal construct validity and cultural equivalence of the SDQ in a New Zealand pre-school population across different ethnicity strata. We hypothesised that the SDQ subscales and the Difficulty scale would i) have cross-informant reliability (with consistency in scores by parents and teachers); ii) fit the Rasch model (demonstrating unidimensionality and internal construct validity), and iii) have cultural equivalence across ethnic strata (demonstrated by an absence of item differential function or DIF).

Methods

Study design and sample

The study utilised SDQ data gathered during the New Zealand Before School Check (B4SC),[3] which takes place when the child is aged 4 or 5. Permission to use the full, de-identified 2011 national SDQ dataset for 4 and 5 year olds from the Ministry of Health was provided by the B4SC Governance Board (n=51,251). Data were included if responses to individual item responses had been entered (as opposed to only total scores). Cases were excluded for District Health Board with fewer than 15% of complete datasets as their data quality was in doubt, and if scores entered were all zero (deemed suspicious as the Prosocial subscale is scored in the opposite direction from the other subscales). In addition, children with ethnicity classed as Other European and Other were excluded as these groups would have contained a very broad number of countries from which children or their families would have hailed, potentially biasing our analysis (especially for differential item functioning by ethnic groups, see below). In total 29,075 cases remained in the parents' dataset (51.3% boys; 68% aged 4; 57% NZ European, 23% Māori, 12% Pasifika 8% Asian); 17,006 remained for the parent-teacher cross-informant reliability analysis.

Fit to the Rasch model is considered acceptable when the observed data fit the predetermined Rasch model,[9 17] traditionally examined with fit statistics (e.g. the item-trait interaction chi-square). A non-significant chi-square indicates fit to the Rasch model. However, power increases with large samples, which inflates the chi-square and results in negligible small differences appearing as a statistically significant misfit between the data and the model.[18 19] Therefore, our analysis was carried out on a smaller sample (n=1,000), to allow examination of convergence to the Rasch model. The sample was created by randomly sampling equal numbers of people for each of the four ethnic groups (250/ethnic group).

Instruments

The SDQ consists of 25 items, each with three response options: not true, somewhat true, and certainly true. The four SDQ subscales reflecting problematic behaviours or emotions (Emotional Symptoms, Conduct Problems, Hyperactivity, Peer Problems) contain 15 positively worded items and five negatively worded items.[1 2] Positively worded items are reverse scored (in New Zealand this is done on data entry), thus higher subscale scores denote greater problems. Scores from these four subscales are also summed to give an overall Difficulty score ranging from 0-40. The five items making up the Prosocial Behaviour subscale are positively worded and higher scores denote better social behaviour.

Data analysis

Cross-informant reliability (between parents and teachers) was assessed for those cases for which both parent and teacher SDQ data were available. The Intraclass Correlation Coefficient (ICC) is the preferred statistical technique and was used.[20 21] However, as many studies of the SDQ have used correlations [22] we will also present those.

Each SDQ subscale and the Difficulty scale were fitted to the Rasch Model to examine fit, using RUMM2030 software.[23] Fit was considered acceptable if there was a non-substantial deviation of individual items and respondents from the Rasch model (individual item and person Fit Residuals should be within the range of +/- 2.5, the average Fit Residual statistics should be close to a mean of zero and standard deviation of one, the item chi-squares should be non-significant). In addition, we used the Root Mean Square Error of Approximation (RMSEA) to examine fit, with RMSEA<0.02 suggesting data fit the Rasch model (Box 1).[19]

Log-transformed item scores generated from the response choices should reflect the increasing or decreasing latent trait to be measured (threshold ordering). When a given level of problems is not confirmed by the expected response option to an item, disordered thresholds are observed. Disordering is only considered statistically significant if the 95% confidence intervals of the threshold locations do not overlap. When significant disordering are observed response categories can be combined.

An assumption of the Rasch model is that the answers to one item should not be dependent on the responses to another item, conditional upon the trait being measured. This local independence is examined by exploring the correlations between items' residuals, which should not be more than 0.20 above the average residual correlation.[24] If locally dependent items are observed they can be combined into a testlet, a bundle of items that share a common stimulus.[25]

The Rasch model expects that each item is invariant (unbiased) across key groups (e.g. ethnicity or gender),[26 27] examined statistically with an Analysis of Variance (ANOVA) and visually by examining the Item Characteristic Curves (ICC). Variance (Differential Item Functioning, DIF) can be uniform; the bias is present consistently across the trait. For example, uniform DIF by ethnic group implies that item difficulty is different for individual ethnic groups across the trait even though their underlying level of problems is the same. DIF can also be non-uniform; the bias is not consistent across the trait. DIF analysis is affected by large sample sizes with non-significant DIF showing as significant, hence inspection of ICCs is also important. When uniform DIF is observed two strategies can be employed. First, DIF items (if present in >1 item) can be combined into a testlet to examine if DIF is cancelled out at the test level; second, the item can be split by the variable for which DIF is observed. In our analysis we considered the final solution to be the one with the best improvements in fit statistics.

Another key assumption of the Rasch model is that a scale must be unidimensional. This is examined by creating two subsets of items, identified by a principal component analysis of the item residuals, with those loading negatively forming one set and those positively loading the second set.[28] An independent t-test is used to compare estimates derived from the two subtests for each respondent. When fewer than 5% of the t-tests are significant (or the 95% confidence interval of t-tests includes 5%) unidimensionality is supported.[28 29]

Targeting of the subscales to the population was examined with person-item-threshold maps.

Internal consistency was examined with Cronbach's Alpha and Person Separation Index (PSI) statistics. PSI is an indicator of the number of statistically different strata (groups) that the test can identify in the sample.[30] Interpretation of the PSI is similar to Cronbach's Alpha with values ≥0.70 suitable for group comparisons and ≥0.85 for individual clinical use. However, Cronbach's Alpha can only be calculated when there are no missing data and are not considered robust with skewed data.[31] Therefore, we present PSI and Cronbach's Alpha in summary tables as well as the number of groups the subscale is able to discriminate between.[32]

Ethical approval was obtained from the New Zealand Health and Disability Ethics Committee (Northern A, NTY/12/04/028/AM05) and the Auckland University of Technology's Ethics Committee (12/163).

Results

Cross-informant reliability

Cross-informant reliability between parent and teachers as measured by correlations was generally poor (all <0.5, mean 0.28) and ICCs (all <0.6, mean 0.13). Cross-informant reliability was better in the Hyperactivity subscale, and worst in the Prosocial subscale; better for NZ European and worst for Pasifika children (table 1).

Internal validity & cross-cultural equivalence

Table 2 displays results from the Rasch analysis.

Emotional Symptoms subscale

All items in this subscale had ordered thresholds, items were locally independent and the subscale was unidimensional. Person fit was adequate with a mean person fit residual reasonably close to 0 and the SD below 1.4 (Table 2: analysis 1). However, overall fit to the Rasch model was unsatisfactory (RMSEA >0.02). PSI was below zero and Cronbach's Alpha (α) 0.15. All item fit residuals were within the acceptable range of -2.5 to 2.5; however, 4 out of 5 item chi-square values were statistically significant, indicating misfit.

There was statistically significant uniform DIF by ethnicity in items 16 and 24, which was confirmed by visual inspection of the ICCs (Figure 1). Items 16 and 24 were combined into a testlet. This resulted in poorer person fit and similar RMSEA values (0.072). We therefore split these items by ethnic groups instead, creating unique items for NZE, Māori, Asian and Pasifika peoples, resulting in 11 items for the subscale. This improved overall fit to the Rasch model, however, the RMSEA was still greater than the acceptable value of 0.02 and internal consistency unacceptably low (Table 2: analysis 2).

After items were split all item fit residuals were within range, although two still had statistically significant chi-square values (items 24NZE and item 8). Table 3 shows that the easiest item to endorse is item 16 and the hardest to endorse is item 13. The split item locations show that for children with the same level of emotional problems item 16 is more readily endorsed when they are Māori and less readily endorsed when they are Pasifika (difference of 0.42 logits). Item 24 is endorsed more readily by parents of Asian than NZE children (difference of 0.49 logits). Figure 2 displays the targeting of the subscale to the population, clearly demonstrating the large number of extreme cases.

Conduct Problems subscale

Conduct Problems item thresholds were ordered, items were locally independent, person fit and unidimensionality were acceptable. However, overall fit to the model was unsatisfactory (RMSEA >0.02, Table 2: analysis 3). Internal consistency was poor (PSI 0.10, α 0.65) with the subscale being able to discriminate between three strata.

Item fit residuals were within acceptable range though two had significant chi-squares (items 5 and 18).

Statistically significant DIF by ethnicity was present for item 12 and by gender for item 7. These two items were split by ethnicity and gender respectively (Table 2: analysis 4), resulting in satisfactory fit residuals, 1 item with a significant chi-square, significant improvement in RMSEA (0.03) but poor internal consistency (PSI=0.11, splitting items leads to missing data and α cannot be calculated). The easiest item to endorse was item 5 and the hardest item 12 (Table 3). The split item locations show that for children with the same level of conduct problems item 12 is more readily endorsed when they are Pasifika and less readily endorsed when they are NZE (difference of 1.22 logits). Item

7 is endorsed more readily by parents of boys than girls (difference of 0.32 logits). Targeting showed a floor effect (Figure 2).

Hyperactivity subscale

Ordered thresholds, local independence, person fit and unidimensionality were observed for the Hyperactivity subscale, however, overall fit to the model and internal consistency was unsatisfactory (RMSE >0.02; PSI 0.30, α 0.48; subscale discriminates between 3 strata, Table 2: analysis 5). Item fit residuals were out of range for item 21 and item 25 had a significant chi-square. Uniform DIF was statistically significant by ethnicity in two items (15 and 21). These items were therefore split by ethnicity. This improved fit to the Rash model (Table 2: analysis 6) and better fit than when these 2 items were combined into a testlet. Item fit residuals were within acceptable range of -2.5/+2.5, only 1 item had a significant item chi-square statistic (Table 3), and RMSEA was close to 0.02. However, internal consistency remained poor (PSI=0.31). The easiest item to endorse was item 15 (for Asian children) and the hardest item 10. The split item locations show that for children with the same level of hyperactivity problems item 15 is more readily endorsed when they are Asian and less readily endorsed when they are NZE (difference of 0.52 logits). Item 21 is endorsed more readily by parents of NZE children than Pasifika children (difference of 0.47 logits, Table 3). The targeting map showed a floor effect (Figure 2).

Peer problems subscale

Ordered thresholds, local independence, person fit and unidimensionality were observed. However, overall fit to the Rasch model and internal consistency were unsatisfactory (RMSEA >0.02; PSI negative value, α 0.51, the subscale is able to discriminate between 2 strata, Table 2: analysis 7). Item fit residuals were acceptable, although two items had significant chi-squares. One item (23) displayed uniform DIF by ethnicity. After splitting this item by ethnicity fit improved; all item fit residuals were within range (item 14 chi-square was borderline statistically significant), RMSEA was close to 0.02. PSI values remained negative, however (Table 2: analysis 8). The easiest item was item 23 (for Asian children) and the hardest item 14. Item 23 was easier for Asian children and hardest for NZE children (difference of 1.10 logits, Table 3). Targeting showed a significant floor effect (Figure 2).

Prosocial subscale

The subscale met the requirements for threshold ordering, local independence, person fit and unidimensionality. Overall fit to the Rasch model and internal consistency were unsatisfactory (RMSEA >0.02; PSI negative values, α 0.29, subscale able to discriminate between 2 strata, Table 2: analysis 9). Item fit residuals were within the -2.5/+2.5 range, tough two had significant item chi-square statistics. There was no DIF. Item 17 was the easiest to endorse; item 4 was the hardest to endorse. A ceiling effect was observed in the person-item-threshold map (Figure 2).

Difficulty scale

Two items had disordered thresholds, however, this was not statistically significant and item response categories did not need to be combined. Some local dependency was present in 2 item pairs. Unidimensionality was observed (Table 2: analysis 10). Five item fit residuals were out of the acceptable range of -2.5/+2.5 and four items showed uniform DIF by ethnicity (items 12, 16, 21 and 23). To examine if DIF was present at the test level these items were combined into a testlet. This resulted in an absence of DIF, however, one item pair remained locally dependent (items 2 and 10). A second testlet was created to deal with this local dependency. The resulting scale was unidimensional, with locally independent items (Table 2: analysis 11). The RMSEA was within range suggesting overall fit to the Rasch model. Internal consistency was good (PSI 0.71, α 0.77, the scale was able to discriminate between 6 distinct strata). The fit residual for one item was slightly out of range (item 15, -2.777), however, given the negative value of this residual this indicates redundancy rather than misfit and the item was therefore retained. The easiest item to endorse was item 15, the hardest item 14. The person-item threshold map shows a normal distribution, although this is located to the left of the item locations on the latent trait. A conversion table was produced, which can be used to convert the raw ordinal score to an interval scale (Table 4).

Discussion

This study has shown that the SDQ items response categories work well, however, the five subscales diverge significantly from the Rasch model and four include items that are biased by key variables (ethnicity having the greatest contribution), raising questions about cultural equivalence. The five subscales suffer from a floor and ceiling effect and their internal consistency statistics are well below the acceptable range. By contrast, the total Difficulty scale, which combines the four subscales capturing children's problems, is unidimensional, fits the Rasch model (after dealing with DIF and local dependency) and has internal consistency sufficient to distinguish between six groups of children. The study has also shown that parents and teachers score children in their care differently. Thus, all three study hypotheses are rejected. This section will discuss our findings in terms of fit to the Rasch model, internal consistency, cultural equivalence and cross-informant reliability.

Fit to the Rasch model

The total Difficulty scale did fit the Rach model, after dealing with four DIF items and two locally dependent items. This scale has good internal consistency and is able to discriminate between six groups of children on the latent trait. We observed the population distribution, whilst following a normal pattern, was to the left of the item locations on the latent trait. Thus, the precision of person estimates at the lower of the scale will not be as good as for those at the higher end of the scale. However, the SDQ is used for screening and arguably precise measurement at the lower end is not needed, since all one needs to establish is that the child does not need to be referred for further assessment or intervention. As we achieved fit to the Rasch model we were able to provide a conversion table which can be used by clinicians to convert the raw ordinal score to more accurate interval level and which takes account of DIF.

Internal consistency

The 5 subscales are relatively short, which affects internal consistency and the subscales' ability to make fine distinctions between groups of people on the underlying trait.[20] In addition, there was significant divergence between the PSI and Cronbach's alpha statistics, with PSI being much smaller than alpha. This divergence can be explained by the way these statistics are calculated. The calculation of Cronbach's alpha assumes all standard errors (SEs) for individuals are the same, making it not a very robust statistics for skewed data.[31] This results in relatively high values even in the presence of extreme scores and the Cronbach alpha values are therefore meaningless for SDQ data. This has not been raised as an issue in the SDQ literature, indeed, Cronbach's Alpha values are widely reported as satisfactory.[33] In Rasch analysis the SE for every individual is estimated and the calculation of the PSI statistic takes these into account. Since SEs are largest for people with extreme scores, PSI will be smaller than the Cronbach's Alpha as observed in our skewed data. However, the purpose of the SDQ is to identify those children who would benefit from further assessment or intervention. Thus, the fact we observed a floor and ceiling effect is not necessarily problematic.

Cultural equivalence

One quantitative study has examined measurement invariance between British Indian and British white children using multi-group confirmatory factor analyses [34] and demonstrated evidence of acceptable fit across ethnicity.[34] However, ours is the first study to examine bias by ethnicity at the item level and found lack of cultural equivalence. DIF (especially by ethnicity) was found for all the four subscales measuring problems, suggesting there are a number of questions to which parents respond differently despite overall scoring the same amount of problems on the trait being measured. If DIF is ignored it could over- or underestimate the child's difficulties since the difficulty of the item varies by ethnic group. Our study is unable to assess why such DIF occurs, since the study drew on secondary data. However, we can pose some possible factors that may have impacted upon this, as discussed below.

Our recent qualitative study suggests there is variation in the way the SDQ is administered – some parents complete the tool by themselves and others receive support from nurses, possibly impacting on the way questions are interpreted.[16] In addition, New Zealand pre-school parents from Māori, Pacific Island, Asian, and new immigrant groups questioned the cultural validity of the SDQ.[16] Respondents in a qualitative study exploring the SDQ in Aboriginal community-controlled health services reported that the use of a questionnaire as opposed to a general conversation or interview was deemed culturally inappropriate and that inter-relationships with peers were considered of less importance than relationships with family and participants.[35]

There are 85 different language versions available from the Youth in Mind website, though not one in Te Reo Māori (http://www.sdqinfo.org/). Translations and adaptations are not permitted without the involvement of that study team, which provides confidence in the robustness of translations. However, for our study we do not know if respondents were offered the SDQ in the language of their choice as such data are not collected as part of the B4SC. The literature includes six studies, which examined SDQ translations, demonstrating some issues with these.[8] Using a language version that is not understood by respondents will affect validity,[36] which may have occurred here.

It is possible that poor literacy impacts on answering the SDQ, as found by others.[37 38] In New Zealand there are many people with poorer than average literacy skills.[39] In addition, 18.6% of the New Zealand population report speaking two or more languages, the majority of these were born overseas (60.4%) and many of these will have English as their second language.[40]

Cross-informant reliability

Cross-informant reliability was examined with ICC's which were well below the acceptable cut off value of 0.6 (the mean in our study was 0.126). However, some argue that correlation coefficients can be used in the assessment of cross-informant reliability of the SDQ since parents and teachers make SDQ ratings based on different sources of information.[1 33] Our systematic literature review found weighted averages of coefficients between different informants ranged from 0.24 to 0.45,[8] similar to findings by others (range 0.26 to 0.47).[33] In our study the mean correlation coefficient was 0.28, meaning only 8% of the variance can be explained by scores from different informants. This implies the importance of taking into account the views of both parents and teachers when making a decision for onward referral, a practice that is not commonplace in New Zealand.[41]

A key strength of this study is the inclusion of all 4 and 5 year old children in New Zealand for whom an SDQ assessment was available in 2011, resulting in our ability to assess the validity of the tool at the population level and with sufficient power to make sounds conclusions. However, we excluded 39% of data as we had some concerns about their quality (it being incomplete or containing multiple inconsistencies). In addition, we were unable to assess DIF by other key variables that may affect validity, e.g. first language or country of birth, as such data were not available.

In conclusion, the total Difficulty scale is internally valid and has acceptable internal consistency. Clinicians should use the conversion table as this takes has taken account of bias by ethnic group. The 5 subscales are not valid and not suitable for use in their own right in New Zealand. Since consistency of scores between parents and teachers was poor it is advisable to use both parents and teachers' feedback when considering children's needs for referral of further assessment. Future work should examine if validity is impacted by different language versions used (in the same country).

Figure 1. Item Characteristics Curves for items from the Strengths and Difficulties Questionnaire (parents, n=1,000)

Figure 2. Person-item-threshold maps Strengths and Difficulties Questionnaire (parents, n=1,000)

Box 1. Calculation of Root Mean Square Error of Approximation (RMSEA)

In Rasch analysis, RMSEA is calculated as follows:

RMSEA = $\sqrt{}$ ([((χ^2/df) - 1)/(N - 1)] , 0) [19]

χ^2 is the item-trait interaction chi-square (obtained from the analysis within the Rasch software),

df is its degrees of freedom

N is the sample size.

Notice that the RMSEA has an expected value of zero when the data fit the model. Overfit of the data to the model, χ^2/df < 1, is ignored. For a given χ^2, RMSEA decreases as sample size (N) increases.

Table 1. Intraclass correlation coefficients SDQ subscales, overall and by ethnicity (n=17,006)

Variable			**Ethnicity**		
	Overall*	**Māori**	**NZ European***	**Pasifika**	**Asian**
	r	**r**	**r**	**r**	**r**
Valid N	17056	2677	10735	1144	1169
Mean item correlations	0.282	0.237	0.315	0.130	0.210
Minimum item correlations	0.199	0.151	0.220	-0.009	0.055
Maximum item correlations	0.418	0.358	0.447	0.275	0.377
	ICC	**ICC**	**ICC**	**ICC**	**ICC**
Emotional Symptoms	0.126	0.067	0.186	0.017	0.098
Conduct Problems	0.137	0.112	0.179	0.038	0.079
Hyperactivity	0.174	0.136	0.245	0.050	0.122
Peer problems	0.139	0.100	0.202	0.004	0.162
Prosocial	0.055	0.048	0.066	0.040	0.035
Mean ICC	0.126	0.093	0.175	0.030	0.099
Minimum ICC	0.055	0.048	0.066	0.004	0.035
Maximum ICC	0.174	0.136	0.245	0.050	0.162

Table 2. Fit to the Rasch model – Strengths and Difficulties Questionnaire parents (SDQ-P) (n=1,000)

Subscales			Item Fit Residual		Person Fit		Chi Square Interaction			RMSEA$^{\%}$	Internal consistency§		Unidimensionality
Analysis name					Residual								T-Tests (CI)$^{\$\$}$
		N	Mean$^{\$}$	SD	Mean	SD	Value	df	P		PSI Without extremes	α Without extremes	% (95% CI)
Emotional Symptoms													
1	Initial	1,000	-0.791	0.894	-0.327	0.783	83.6	20	<0.0001	0.068	-0.40	0.15	0
2	Split items 16&24	1,000	-0.545	0.841	-0.343	0.735	99.1	41	<0.0001	0.045	-0.41	N/A	0
Conduct Problems													
3	Initial	1,000	0.266	1.273	-0.253	0.876	71.6	20	<0.0001	0.060	0.10	0.65	0
4	Split items 7&12	1,000	0.134	0.902	-0.254	0.882	75.3	45	0.003	0.031	0.11	-	0
Hyper-activity													
5	Initial	1,000	0.260	2.348	-0.359	1.147	97.3	25	<0.0001	0.06	0.30	0.48	0.5 (-1.0 to 2.0)
6	Split items 15&21	1,000	0.323	1.480	-0.365	1.134	125.6	69	<0.0001	0.03	0.31	-	0.5 (-1.0 to 2.0)
Peer Problems													

7	Initial	1,000	-0.339	0.868	-0.207	0.719	69.0	20	<0.0001	0.06	-0.49	0.51	0
8	Split item 23	1,000	-0.207	0.652	-0.213	0.733	79.5	52	0.008	0.03	-0.43	-	0
Prosocial													
9	Initial	1,000	-0.075	1.592	-0.319	1.079	66.6	20	<0.0001	0.06	-0.03	0.29	0.1 (-1.5 to 1.8)
Difficulty													
10	Initial	1,000	-0.448	1.848	-0.248	1.004	296.3	180	0.0001	0.03	0.71	0.79	5.9 (4.6 to 7.3)
11	Testlets DIF$^{\%\%}$ items & LD§§ items	1,000	-0.615	1.321	-0.294	0.985	200.4	144	0.001	0.02	0.71	0.77	3.0 (1.6 to 4.4)

***Note* - Indices indicative of fit:**

[$] Mean item and person fit residuals: should be close to 0 (and <0.4); SD close to 1 (and <1.4)

[%] RMSEA (Root Mean Square Error of Approximation) <0.02

[§] Internal consistency PSI and α ≥0.70 (allows for group comparisons) and ≥0.85 (allows for individual clinical use)

[$$] Unidimensionality indicated if fewer than 5% of t-tests are significant (i.e. the 95% CI should include 5%)

[%%] DIF: Differential item Functioning

[§§] LD: Local Dependency

Table 3. Item locations (in location order) and fit statistics Strengths and Difficulties Questionnaire parents (SDQ-P) subscales (n=1,000)

Subscale & Items	**Location**	**SE**	**Fit Residual**	**Chi Square value**	**df**	**P**
Emotional problems [$]						
16 Māori	-0.871	0.113	-0.226	2.968	4	0.5631
16 NZE	-0.692	0.124	-0.036	0.60	3	0.8960
16 Asian	-0.538	0.118	-0.101	0.77	3	0.8569
16 Pasifika	-0.450	0.120	0.911	0.61	3	0.8936
24 Asian	-0.250	0.124	-0.185	5.13	3	0.1629
24 Māori	0.010	0.117	-0.737	9.69	4	0.0461
24 Pasifika	0.024	0.124	-0.002	11.857	3	0.0079
24 NZE	0.243	0.127	-1.610	14.095	3	0.0028
3	0.653	0.070	-0.615	15.156	5	0.0097
8	0.908	0.075	-1.970	21.479	5	0.0007
13	0.965	0.080	-1.423	16.749	5	0.0050
Conduct Problems [%]						
5	-0.985	0.063	0.011	15.38	5	0.0089
18	-0.707	0.066	-0.352	22.19	5	0.0005
7 Male	-0.594	0.096	1.209	7.71	5	0.1732
7 Fem	-0.271	0.100	1.917	6.09	5	0.2975
22	-0.012	0.072	0.156	8.49	5	0.1312
12 Pasifika	0.089	0.143	-0.148	3.527	5	0.6193
12 Māori	0.339	0.145	-0.512	5.862	5	0.3199
12 Asian	0.838	0.202	-0.030	2.344	5	0.7998
12 NZE	1.304	0.211	-1.049	3.733	5	0.5884
Hyperactivity [$]						
15Asian	-0.491	0.109	-0.395	8.25	5	0.1432
15 Māori	-0.315	0.117	0.433	1.78	6	0.9388

21 NZE	-0.234	0.142	2.204	17.50	5	0.0037
2	-0.206	0.056	-1.327	23.29	9	0.0056
21 Asian	-0.186	0.124	1.414	8.216	5	0.1447
15 Pasifika	-0.019	0.121	0.388	8.775	5	0.1184
15 NZE	0.032	0.126	-1.737	12.772	5	0.0256
21 Māori	0.114	0.129	1.743	7.403	6	0.2852
21 Pasifika	0.234	0.122	1.393	5.986	5	0.3076
25	0.360	0.066	1.421	9.335	9	0.4070
10	0.712	0.065	-1.984	22.26	9	0.0081
Peer Problems [%]						
23 A	-0.968	0.109	-0.571	1.959	4	0.7432
23 P	-0.870	0.107	0.307	4.311	5	0.5056
23 M	-0.217	0.119	0.038	5.529	4	0.2372
6	-0.026	0.065	0.526	10.572	9	0.3062
23 N	0.130	0.154	0.093	3.548	3	0.3147
11	0.233	0.066	-1.419	17.787	9	0.0377
19	0.491	0.071	0.131	12.305	9	0.1967
14	1.227	0.084	-0.763	23.501	9	0.0052
Prosocial [§]						
1	-0.487	0.079	-1.530	18.205	4	0.0011
4	-0.036	0.073	-0.273	12.624	4	0.0133
9	0.000	0.072	1.092	6.74	4	0.1502
17	0.008	0.071	-1.633	21.52	4	0.0003
20	0.515	0.073	1.972	7.52	4	0.1109
Difficulty [$$]						
15	-0.835	0.054	-2.777	27.39	9	0.0012
LD items [%%]	-0.606	0.037	-1.744	14.01	9	0.1221
5	-0.583	0.056	-0.595	8.71	9	0.4645
DIF items [§§]	-0.375	0.031	-2.500	21.03	9	0.0125
25	-0.331	0.061	0.036	14.05	9	0.1207

24	-0.314	0.058	0.839	7.44	9	0.5911
18	-0.313	0.059	-0.742	6.83	9	0.6553
6	-0.137	0.061	1.137	4.47	9	0.8777
7	-0.026	0.063	-1.305	23.26	9	0.0057
11	0.117	0.067	0.862	9.76	9	0.3702
22	0.308	0.068	-1.218	14.07	9	0.1199
3	0.311	0.071	1.017	11.50	9	0.2433
19	0.413	0.072	-1.247	10.59	9	0.3048
8	0.561	0.077	0.105	4.79	9	0.8525
13	0.646	0.087	0.621	9.37	9	0.4035
14	1.164	0.084	-2.326	13.15	9	0.1560

Note

[$] Bonferroni corrections applied P is statistically significant if <0.005

[%] Bonferroni corrections applied P is statistically significant if <0.006

[§] Bonferroni corrections applied P is statistically significant if <0.01

[$$] Bonferroni corrections applied P is statistically significant if <0.003

[%%] LD (Locally Dependent) items; combined into a testlet (item 2 and 10)

[§§] DIF (Differential Item Functioning) items combined into a testlet (items 12, 16, 21, 23)

Table 4. Conversion table for the Difficulty scale of the Strengths and Difficulties Questionnaire parents (SDQ-P)

Original Total Difficulty score (ordinal data)	Logit scores (interval level data)	Converted logit scores to 0-40 scale (interval level data)
0	-4.483	0
1	-3.655	4
2	-3.082	7
3	-2.685	8
4	-2.375	10
5	-2.117	11
6	-1.895	12
7	-1.699	13
8	-1.522	14
9	-1.36	15
10	-1.209	15
11	-1.068	16
12	-0.935	16
13	-0.809	17
14	-0.687	18
15	-0.571	18
16	-0.457	19
17	-0.347	19
18	-0.24	20
19	-0.134	20
20	-0.029	21
21	0.075	21
22	0.178	22
23	0.282	22

24	0.386	23
25	0.492	23
26	0.599	24
27	0.709	24
28	0.822	25
29	0.94	25
30	1.064	26
31	1.196	26
32	1.337	27
33	1.491	28
34	1.663	29
35	1.859	29
36	2.09	31
37	2.373	32
38	2.746	34
39	3.301	36
40	4.125	40

Author contributions

PK conceived of the study, led on study design, project management, data analysis and dissemination. AV, HE, KMcP contributed to study design. AV contributed to the data analysis. PK drafted the manuscript and is the guarantor. All authors revised it critically for important intellectual content and approved the final version for publication. All authors agree to be accountable for all aspects of the work.

Competing Interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: PK, AV, HE, KMcP had financial support from the Ministry of Health of New Zealand for the submitted work; subsequent to the completion of this project and data analysis KMcP became the Chief Executive of the Health Research Council of New Zealand; all other authors declare no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the funder.

Data sharing

Quantitative data from the study can be obtained from the author, subject to the funder's permission.

References

1. Goodman R. The strengths and difficulties questionnaire: A research note. Journal of Child Psychology and Psychiatry and Allied Disciplines 1997;**38**(5):581-86.

2. Goodman R, Meltzer H, Bailey V. The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. European Child & Adolescent Psychiatry 1998;**7**(3):125-30.

3. Ministry of Health. The B4 School Check. A handbook for practitioners. Wellington, 2008.

4. Williamson A, Este C, Clapham K, et al. What are the factors associated with good mental health among Aboriginal children in urban New South Wales, Australia? Phase I findings from the Study of Environment on Aboriginal Resilience and Child Health (SEARCH). BMJ Open 2016;**6**(7).

5. Klein AM, Otto Y, Fuchs S, et al. Psychometric properties of the parent-rated SDQ in preschoolers. European Journal of Psychological Assessment 2013;**29**(2):96-104.

6. Tobia V, Gabriele MA, Marzocchi GM. The Italian Version of the Strengths and Difficulties Questionnaire (SDQ)-Teacher: Psychometric Properties. Journal of Psychoeducational Assessment 2013;**31**(5):493-505.

7. Mieloo CL, Bevaart F, Donker MC, et al. Validation of the SDQ in a multi-ethnic population of young children. The European Journal of Public Health 2014;**24**(1):26-32.

8. Kersten P, Czuba K, McPherson KM, et al. A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. International Journal of Behavioral Development 2016;**40**(1):64-75.

9. Rasch G. *Probabilistic models for some intelligence and attainment tests (revised and expanded ed.)*. Chicago: The University of Chicago Press, 1960/1980.

10. Wright BD. Comparing Rasch measurement and factor analysis. Structural Equation Modeling 1996;**3**:3-24.

11. Christensen KB, Engelhard J, G., Salzberger T. Rasch vs. Factor Analysis. Rasch Measurement Transactions 2012;**26**(3):1373-8.

12. Borg A-M, Pälvi K, Raili S, et al. Reliability of the Strengths and Difficulties Questionnaire among Finnish 4-9-year-old children. Nordic Journal of Psychiatry 2012;**66**(6):403-13.

13. Statistics New Zealand. Māori Population Estimates: Mean year ended 31 December 2016. http://archive.stats.govt.nz/browse_for_stats/population/estimates_and_projections/MaoriPopulationEstimates_HOTPMYe31Dec16.aspx. Downloaded 4 Jan 2018.

14. Høegh MC, Høegh S-M. Trans-adapting outcome measures in rehabilitation: Cross-cultural issues. Neuropsychological Rehabilitation 2009;**19**(6):955-70.

15. de Klerk G. Cross-Cultural Testing In: Born M, Foxcroft, C.D., Butter, R., ed. Online Readings in Testing and Assessment: International Test Commission, 2008.

16. Kersten P, Dudley M, Nayar S, et al. Cross-cultural acceptability and utility of the strengths and difficulties questionnaire: views of families. BMC Psychiatry 2016;**16**(1):347.

17. Andrich D. *Rasch models for measurement series: quantitative applications in the social sciences no. 68*. London: Sage Publications, 1988.

18. Linacre JM. Rasch Power Analysis: Size vs. Significance: Infit and Outfit Mean-Square and Standardized Chi-Square Fit Statistic. Rasch Meas Trans 2003;**17**(1):918.

19. Tennant A, Pallant JF. The Root Mean Square Error of Approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. Rasch Meas Trans 2012;**25**(4):1348-9.

20. Streiner DL, Norman GR. *Health Measurement Scales: a practical guide to their development and use*. Oxford: Oxford University Press, 2008.

21. Charter RA, Feldt LS. Confidence intervals for true scores: Is there a correct approach? Journal of Psychoeducational Assessment 2001;**19**(4):350-64.

22. Nunnally JC, Bernstein IH. *Psychometric theory (3rd ed.)*. New York: McGraw-Hill, 1994.

23. RUMM2030 [program]: RUMM Laboratory Pty Ltd., 2009.

24. Marais I, Andrich D. Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. Journal of Applied Measurement 2008;**9**(2):105-24.

25. Wainer H, Kiely G. Item clusters and computer adaptive testing: A case for testlets. J Educ measurement 1987;**24**:185-202.

26. Grimby G. Useful reporting of DIF. Rasch Measurement Transactions 1998;**12**(3):651.

27. Holland PW, Wainer H. Differential Item Functioning. NJ: Hillsdale. Lawrence Erlbaum, 1993.

28. Smith EV. Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement 2002;**3**:205-31.

29. Tennant A, Pallant JF. Unidimensionality matters! (a tale of two Smiths?). Rasch Measurement Transactions 2006;**20**:1048-51.

30. Wright BD. Reliability and separation. Rasch Measurement Transactions 1996;**9**(4):472.

31. Sheng Y, Sheng Z. Is Coefficient Alpha Robust to Non-Normal Data? Frontiers in Psychology 2012;**34**(1):1-13.

32. Wright BD. Separation, Reliability and Skewed Distributions: Statistically Different Levels of Performance. Rasch Meas Trans 2001;**14**(4):786.

33. Stone LL, Otten R, Engels RC, et al. Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-year-olds: a review. Clinical Child And Family Psychology Review 2010;**13**(3):254-74.

34. Goodman A, Patel V, Leon DA. Why do British Indian children have an apparent mental health advantage? Journal of Child Psychology and Psychiatry and Allied Disciplines 2010;**51**(10):1171-83.

35. Williamson A, Redman S, Dadds M, et al. Acceptability of an emotional and behavioural screening tool for children in Aboriginal Community Controlled Health Services in urban NSW. Australian and New Zealand Journal of Psychiatry 2010;**44**(10):894-900.

36. Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;**25**(24):3186-91.

37. Samad L, Hollis C, Prince M, et al. Child and adolescent psychopathology in a developing country: Testing the validity of the Strengths and Difficulties Questionnaire (Urdu version). International Journal Of Methods In Psychiatric Research 2005;**14**(3):158-66.

38. Thabet AA, Stretch D, Vostanis P. Child mental health problems in Arab children: Application of the strengths and difficulties questionnaire. International Journal of Social Psychiatry 2000;**46**(4):266-80.

39. Lane C. Adult literacy and numeracy in New Zealand – A regional analysis. Perspectives from the Adult Literacy and Life Skills Survey, 2012.

40. Statistics New Zealand. 2013 Census QuickStats about culture and identity. Available from http://www.stats.govt.nz. Wellington, New Zealand, 2014.

41. Hedley C, Thompson S, Morris Mathews K, et al. The B4 School Check behaviour measures: Findings from the hawke's bay evaluation. Nursing Praxis in New Zealand 2012;**28**(3):13-23.

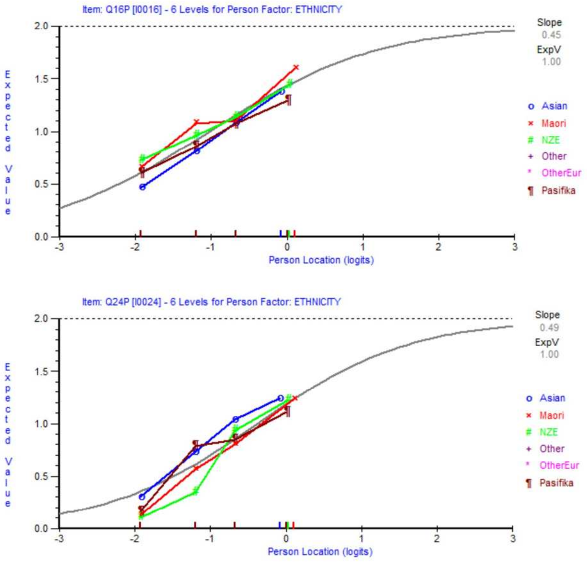
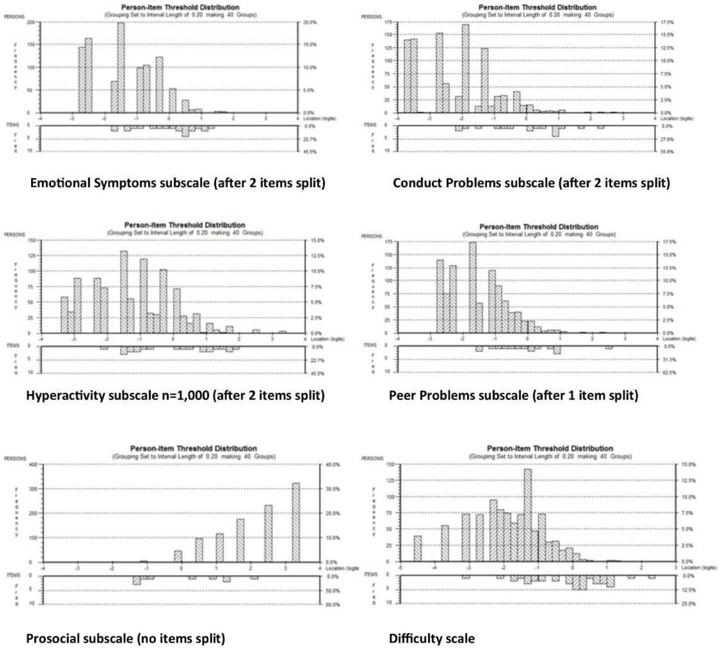


Figure 1. Item Characteristics Curves for items from the Strengths and Difficulties Questionnaire (parents, n=1,000)

209x297mm (300 x 300 DPI)



Emotional Symptoms subscale (after 2 items split)

Conduct Problems subscale (after 2 items split)

Hyperactivity subscale n=1,000 (after 2 items split)

Peer Problems subscale (after 1 item split)

Prosocial subscale (no items split)

Difficulty scale

Figure 2. Person-item-threshold maps Strengths and Difficulties Questionnaire (parents, n=1,000)

209x297mm (300 x 300 DPI)

BMJ Open

Strengths and Difficulties Questionnaire: internal validity and reliability for New Zealand pre-schoolers

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2018-021551.R1
Article Type:	Research
Date Submitted by the Author:	07-Mar-2018
Complete List of Authors:	Kersten, Paula; University of Brighton, School of Health Sciences Vandal, Alain; AUT University, Department of Biostatistics and Epidemiology; Counties Manukau District Health Board, Ko Awatea Health Intelligence and Informatics Elder, Hinemoa; Te Whare Wānanga o Awanuiārangi, School of Graduate Studies McPherson, Kathryn; Health Research Council of New Zealand; AUT University, Centre for Person Centred Research, School of Clinical Sciences
<b>Primary Subject Heading</b>:	Research methods
Secondary Subject Heading:	Public health, Research methods
Keywords:	Strengths and Difficulties Questionnaire, validity, reliability, Rasch, pre-school

Strengths and Difficulties Questionnaire: internal validity and reliability for New Zealand pre-schoolers

Paula Kersten[1], Alain C Vandal[2], Hinemoa Elder[3], Kathryn M McPherson[4 5]

[1] School of Health Sciences, University of Brighton, UK

[2] Department of Biostatistics and Epidemiology, AUT University, New Zealand & Ko Awatea Health Intelligence and Informatics, Counties Manukau District Health Board, New Zealand

[3] Te Whare Wānanga o Awanuiārangi, Auckland, New Zealand.

[4] The Health Research Council of New Zealand

[5] Centre for Person Centred Research, School of Clinical Sciences, AUT University, New Zealand.

Corresponding author

Professor Paula Kersten, School of Health Sciences, University of Brighton, Westlain House, Falmer, Brighton BN1 9PH. p.kersten@brighton.ac.uk. Tel +44 1273 643483. Fax: +44 1273 644010.

Abstract

Objectives: This observational study examines the internal construct validity, internal consistency and cross-informant reliability of the Strengths and Difficulties Questionnaire (SDQ) in a New Zealand pre-school population across four ethnicity strata (New Zealand European, Māori, Pasifika, Asian).

Design: Rasch analysis was employed to examine internal validity on a subsample of 1,000 children. Internal consistency (n=29,075) and cross-informant reliability (n=17,006) was examined using correlations, intraclass correlation coefficients and Cronbach's Alpha on the sample available for such analyses.

Setting & participants: Data were utilised from a national SDQ database provided by the funder, pertaining to New Zealand domiciled children aged 4 and 5, and scored by their parents and teachers.

Results: The five subscales do not fit the Rasch model (as indicated by the overall fit statistics); contain items that are biased (differential item functioning) by key variables, suffer from a floor and ceiling effect and have unacceptable internal consistency. After dealing with differential item functioning the Total Difficulty scale does fit the Rasch model and has good internal consistency. Parent/teacher inter-rater reliability was unacceptably low for all subscales.

Conclusion: The five SDQ subscales are not valid and not suitable for use in their own right in New Zealand. We have provided a conversion table for the Total Difficulty scale, which takes account of bias by ethnic group. Clinicians should use this conversion table in order to reconcile differential item functioning by culture in final scores. It is advisable to use both parents and teachers' feedback when considering children's needs for referral of further assessment. Future work should examine whether validity is impacted by different language versions used in the same country.

Keywords

Strengths and Difficulties Questionnaire, validity, reliability, Rasch, pre-school

Strengths and limitations of this study

- A key strength of this study is the inclusion of all 4 and 5 year old children in New Zealand for whom an SDQ assessment was available in 2011, resulting in our ability to assess the validity of the tool at the population level and with sufficient power to make sound conclusions.
- A strength of the study included robust data quality checks, and the exclusion of 39% of cases for which we had concerns about their quality (it being incomplete or containing multiple inconsistencies).
- A limitation was our inability to assess DIF by other key variables that may affect validity, e.g. first language or country of birth, as such data were not available.
- Future work should examine whether validity is impacted by different language versions used (in the same country).

Funding

This work was supported by the Ministry of Health of New Zealand (grant number 341088). The funding body has not had input into the design, data collection, analysis, interpretation of data, in the writing of the manuscript, nor in the decision to submit the manuscript for publication.

Acknowledgements

We thank the funder for supporting the study.

Introduction

Educational achievement and problems in primary and secondary school aged children can arise as a result of behavioural and emotional problems when the child is of pre-school age.[1-5] Consequently, screening to identify children with, or at risk of behavioural problems at a pre-school age is an increasingly used preventative strategy, aiming to enhance the success of support programmes and early intervention.[6] Such screening is best performed using standardised methods, and for behavioural assessment this means the use of a questionnaire based measure. The Strengths and Difficulties Questionnaire for parents (SDQ-P) and for teachers (SDQ-T) is a tool used worldwide for this purpose to screen pre-school children's psychosocial attributes (positive and negative behaviours).[7-10] It consists of 25 items, making up five subscales: Emotional Symptoms, Conduct Problems, Hyperactivity, Peer Problems, and Prosocial Behaviour.[7 8]

Before using a measure such as the SDQ, establishing validity and reliability is key for optimum decision-making. At present there are two dominant approaches to the development and testing of measures: Classical Test Theory (CTT) and Modern Test Theory (also known as item response theory).[11] In CTT it is assumed that the observed scores on items are the sum of the true score (which we cannot directly measure) and measurement error. However, neither the true score, nor the measurement error can be determined and the approach is therefore flawed.[12] In addition, the best conclusion that can be made following satisfactory tests of validity and reliability using CTT is that an outcome measure is an ordinal scale. Yet, many statistical tests that examine the validity of scales assume that the data arising are of interval nature. Indeed, in the pre-school population, the SDQ has only been tested using parametric, CTT approaches, as demonstrated in our recent systematic review[13] to which we return below. By contrast, Modern Test Theory approaches, such as Rasch analysis, are underpinned by mathematical models that specify the conditions under which equal interval measurements can be estimated from outcome measurement data.[14-16] These approaches are therefore more robust.

Evaluations of the structural validity of the SDQ drawing on CTT in pre-schoolers has been extensively researched using factor analysis (e.g. by [17-19]) , Cronbach's alphas (α)[13] and correlation coefficients,[13 20] and Weighted Least Squares in older children.[21] Our systematic review found acceptable to good evidence for the 5-factor SDQ structure in pre-schoolers, when confirmatory factor analysis (CFA) had been used.[13] A different approach to examining structural validity, using Modern Test Theory, can be achieved by examining whether each of the subscales are unidimensional and fit the Rasch model (i.e. examining internal construct validity).[15] Like CFA, Rasch analysis is a confirmatory approach to examining whether items belong to the subscales under investigation. However, there are known limitations to using factor analysis on ordinal scales, including its parametric basis and the emergence of 'difficulty factors', which may spuriously indicate multidimensionality.[22] In addition, factor analysis does not allow detailed investigation of item function in regard to targeting, differential item functioning and local dependency between items, whereas Rasch analysis includes such assessments.[23] We identified one study which had employed Rasch analysis on SDQ data that had been self-completed by 12 to 18 year olds in Sweden.[24] This study showed that none of the SDQ scales was psychometrically robust, with mis-fitting items in all five subscales and poor internal consistency. However, that study did not examine whether the scale was invariant across different subgroups.

Internal consistency of the SDQ-P subscales has been reported in many studies and synthesised in a systematic review.[13] The sample size-weighted average Cronbach's alphas (α) for the five subscales was below the threshold of 0.70 (implying inadequate internal consistency for shorter, established scales); and for the Difficulty scale α was 0.79 (acceptable for group comparisons but not for individual use). [25 (p91)]

Inter-rater reliability of SDQ subscales between two parents and between two teachers has previously been found to be acceptable when correlation coefficients were used (between 0.42 and 0.64 for parents and between 0.59 and 0.81 for teachers.[20] Other studies have examined scores between different types of informants (e.g. parent and teacher). The systematic review showed that the sample size-weighted average correlation coefficients generated from these studies were weak to moderate (between 0.25 and 0.45).[13]

The validity and reliability of the SDQ have not previously been examined in New Zealand, a country with a sizeable indigenous population (Māori, 15.4%) and immigrant population (25.2% born overseas).[26] New Zealand is a multi-cultural society, impacting upon values, ways of living and languages spoken. It cannot be assumed that measures capturing psychological constructs will have cultural equivalence.[27 28] Indeed, a New Zealand qualitative study has shown that parents from Māori, Pacific Island, Asian, and new immigrant groups questioned the cultural validity of the SDQ.[29] Cultural equivalence therefore needs further investigation.

In summary, the use of Classical Test Theory approaches to examine the validity of the SDQ are limited, evidence suggests cross-informant reliability is weak, and there is no evidence for cultural equivalence for the New Zealand population. Therefore, we aimed to use Modern Test Theory, and specifically Rasch analysis, to examine the internal construct validity and cultural equivalence of the SDQ in a New Zealand pre-school population across different ethnicity strata; and to examine reliability between parents and teachers (cross-informant reliability). We hypothesised that the SDQ subscales and the Difficulty scale would i) have cross-informant reliability (with consistency in scores by parents and teachers); ii) fit the Rasch model (demonstrating unidimensionality and internal construct validity), and iii) have cultural equivalence across ethnic strata (demonstrated by an absence of item differential function or DIF).

Methods

Study design and sample

This observational study utilised SDQ data gathered during the New Zealand Before School Check (B4SC), which takes place when the child is aged (4 or exceptionally aged 5).[9] The B4SC is carried out by registered nurses based in primary care and involves the assessment of the child's general health, hearing, oral health, vision, growth as well as developmental and behavioural problems. The latter is evaluated using the Australian SDQ version for 2 to 4 year olds, completed by the parent. If the child is in pre-school the nurse also requests their teacher to complete the SDQ for the child. Clear instructions for the administration of the SDQ are provided within the B4SC handbook. In New Zealand there is no other SDQ data collection point during childhood.

Data sources/quality, missing data and bias: Permission to use the full, de-identified 2011 national B4SC SDQ dataset for pre-schoolers (n=51,251) from the New Zealand Ministry of Health was provided by the B4SC Governance Board. Data quality checks on SDQ data resulted in the deletion of 20,024 cases (out of n=51,251, 39%) for the following reasons:

1. Individual item data from the parent questionnaire were missing completely (n=19,197) or partially (n=1) since a) we would not have been able to carry out a quality check of the subscale scores, and b) we would not be able to use these data for the Rasch analysis); thus 19,198 were removed from the analysis set;
2. District Health Boards (DHB) for which we had fewer than 15% of data on individual items, since the quality of their data is in doubt: although a total of 12,720 records came from these DHBs, this extra step only entailed the removal of a further 375 records from the analysis set after step 1;
3. Children's ages were recorded as younger than 4 or older than 5 when the SDQ was completed (we suspect some of these ages may have been entered incorrectly; however this step only entailed the removal of a further 451 records from the analysis set after steps 1 and 2;
4. Cases with all zero scores: these were deemed potentially erroneous as the Prosocial subscale is scored in the opposite direction from the other subscales; although 1,038 cases fitted this profile, none had complete parental item data, and so no further record was removed on the basis of this criterion after steps 1, 2 and 3.

Study size: In total 29,075 cases remained in the parents' dataset; 17,006 remained for the parent-teacher cross-informant reliability analysis. Rasch analysis uses fit statistics, but these are not suited to such large sample sizes. Fit to the Rasch model is considered acceptable when the observed data fit the predetermined Rasch model,[15 30] traditionally examined with fit statistics (e.g. the item-trait interaction chi-square). A non-significant chi-square indicates fit to the Rasch model. Power increases with large samples, which inflates the chi-square and results in negligible small differences appearing as a statistically significant misfit between the data and the model.[31 32] Therefore, our Rasch analysis was carried out on a smaller sample (n=1,000), to allow examination of convergence to the Rasch model. The sample was created by randomly sampling equal numbers of cases from the total parent sample, for four main ethnic groups (250/ethnic group): New Zealand European (NZE), Māori, Asian and Pasifika. This is well above the recommended sample size for studies using Rasch analysis. For example, it has been suggested that to have 99% confidence that the estimated item difficulty is within +/- ½ logit of its stable value on the interval metric, the minimum sample size range is 108 to 243 (best to poor targeting).[33 34]

Instruments

The SDQ consists of 25 items, each with three response options: not true, somewhat true, and certainly true. The four SDQ subscales reflecting problematic behaviours or emotions (Emotional Symptoms, Conduct Problems, Hyperactivity, Peer Problems) contain 15 positively worded items and five negatively worded items.[7 8] Positively worded items are reverse scored (in New Zealand this is done on data entry), thus higher subscale scores denote greater problems. Scores from these four subscales are also summed to give an overall Difficulty score ranging from 0-40. The five items making up the Prosocial Behaviour subscale are positively worded and higher scores denote better social behaviour.

Data analysis

Cross-informant reliability (between parents and teachers) was assessed for those cases for which both parent and teacher SDQ data were available (n=17,006). The Intraclass Correlation Coefficient (ICC) is the preferred statistical technique and was used.[25 35] However, as many studies of the SDQ have used correlations [36] we will also present those.

Each SDQ subscale and the Difficulty scale were fitted to the Rasch model to examine fit, using RUMM2030 software.[37] Fit was considered acceptable if there was a non-substantial deviation of individual items and respondents from the Rasch model (individual item and person Fit Residuals should be within the range of +/- 2.5, the average Fit Residual statistics should be close to a mean of zero and standard deviation of one, the item chi-squares should be non-significant). In addition, we used the Root Mean Square Error of Approximation (RMSEA) to examine fit, with RMSEA<0.02 suggesting data fit the Rasch model (Box 1).[32]

Log-transformed item scores generated from the response choices should reflect the increasing or decreasing latent trait to be measured (threshold ordering).[30] When a given level of problems is not confirmed by the expected response option to an item, disordered thresholds are observed. Disordering is only considered statistically significant if the 95% confidence intervals of the threshold locations do not overlap. When significant disordering is observed response categories can be combined.

An assumption of the Rasch model is that the answers to one item should not be dependent on the responses to another item, conditional upon the trait being measured. This local independence is examined by exploring the correlations between items' residuals, which should not be more than 0.20 above the average residual correlation.[38] If locally dependent items are observed they can be combined into a testlet, a bundle of items that share a common stimulus.[39]

The Rasch model expects that each item is invariant (unbiased) across key groups (e.g. ethnicity or gender),[40 41] examined statistically with an Analysis of Variance (ANOVA) and visually by examining the Item Characteristic Curves (ICC). Variance (Differential Item Functioning, DIF) can be uniform; the bias is present consistently across the trait. For example, uniform DIF by ethnic group implies that item difficulty is different for individual ethnic groups across the trait even though their underlying level of problems is the same. DIF can also be non-uniform; the bias is not consistent across the trait.

DIF analysis is affected by large sample sizes with non-significant DIF showing as significant; hence inspection of ICCs is also important. When uniform DIF is observed two strategies can be employed. First, DIF items (if present in >1 item) can be combined into a testlet to examine if DIF is cancelled out at the test level; second, the item can be split by the variable for which DIF is observed. In our analysis we considered the final solution to be the one with the best improvements in fit statistics.

Another key assumption of the Rasch model is that a scale must be unidimensional. This is examined by creating two subsets of items, identified by a principal component analysis of the item residuals, with those loading negatively forming one set and those positively loading the second set.[42] An independent t-test is used to compare estimates derived from the two subtests for each respondent. When fewer than 5% of the t-tests are significant (or the 95% confidence interval of t-tests includes 5%) unidimensionality is supported.[42 43]

Targeting of the subscales to the population was examined with person-item-threshold maps.

Internal consistency was examined with Cronbach's alpha and Person Separation Index (PSI) statistics. PSI is an indicator of the number of statistically different strata (groups) that the test can identify in the sample.[44] Interpretation of the PSI is similar to Cronbach's alpha with values ≥0.70 suitable for group comparisons and ≥0.85 for individual clinical use. However, Cronbach's alpha can only be calculated when there are no missing data and is not considered robust with skewed data.[45] Therefore, we present PSI and Cronbach's alpha in summary tables as well as the number of groups between which the subscale is able to discriminate.[46]

Finally, for polytomous scales two Rasch models can be used. The Rating Scale version assumes that the distance between thresholds is equal across items.[47] The Unrestricted (Partial Credit) model does not make this assumption.[48] A log-likelihood test examines whether results from these two models are significantly different and if this is so the Partial Credit model should be used. This test was significant ($p<0.001$) for all subscales and therefore the Partial Credit model was used.

Patient and Public Involvement

End users of our research include families, pre-school teachers, service providers and the Ministry of Health. The research aims and questions were part of a tender prepared by the Ministry of Health, to which we responded. Thus, we did not have the ability to include end users in the development of study questions. The analysis presented here did not require participant recruitment or data collection and end users were therefore not consulted about the study design. Researchers in New Zealand have a responsibility to ensure their research is of value and culturally responsive to Māori. Therefore, guidance for the study was sought from the University's Mātauranga Māori committee, which members are drawn from a wide range of Māori communities. The findings from the part of the study reported here were presented to the Ministry of Health.

Ethical approval was obtained from the New Zealand Health and Disability Ethics Committee (Northern A, NTY/12/04/028/AM05) and the Auckland University of Technology's Ethics Committee (12/163).

Results

The child gender split was balanced with 49% female and 51% male in the full parent sample, as well as the cross-comparison sample; 99.6% were aged four at the time of the B4SC (0.4% of children had recently turned 5). Child ethnicity in the parent sample was 57% NZE, 23% Māori, 12% Pasifika, and 8% Asian; this distribution was similar in the cross-comparison sample 63% NZE, 16% Māori, 7% Pasifika, and 7% Asian. As noted above, there were no missing data in the selected samples.

Cross-informant reliability (n=17,006)

Cross-informant reliability between parent and teachers as measured by correlations was generally poor (all <0.5, mean 0.28) and ICCs (all <0.6, mean 0.13). Cross-informant reliability was better in the Hyperactivity subscale, and worst in the Prosocial subscale; better for NZE and worst for Pasifika children (table 1).

Internal validity & cross-cultural equivalence

Table 2 displays results from the Rasch analysis.

Emotional Symptoms subscale

All items in this subscale had ordered thresholds, items were locally independent and the subscale was unidimensional. Person fit was adequate with a mean person fit residual reasonably close to 0 and the SD below 1.4 (Table 2: analysis 1). However, overall fit to the Rasch model was unsatisfactory (RMSEA >0.02). PSI was below zero and Cronbach's alpha (α) 0.15. All item fit residuals were within the acceptable range of -2.5 to 2.5; however, four out of five item chi-square values were statistically significant, indicating misfit.

There was statistically significant uniform DIF by ethnicity in items 16 and 24, which was confirmed by visual inspection of the ICCs (figure 1). Items 16 and 24 were combined into a testlet. This resulted in poorer person fit and similar RMSEA values (0.072). We therefore split these items by ethnic groups instead, creating unique items for NZE, Māori, Asian and Pasifika peoples, resulting in 11 items for the subscale. This step improved overall fit to the Rasch model, however, the RMSEA was still greater than the acceptable value of 0.02 and internal consistency unacceptably low (Table 2: analysis 2).

After items were split all item fit residuals were within range, although two still had statistically significant chi-square values (items 24NZE and item 8). Table 3 shows that the easiest item to endorse is item 16 and the hardest to endorse is item 13. The split item locations show that for children with the same level of emotional problems item 16 is more readily endorsed when they are Māori and less readily endorsed when they are Pasifika (difference of 0.42 logits). Item 24 is endorsed more readily by parents of Asian than NZE children (difference of 0.49 logits). Figure 2 displays the targeting of the subscale to the population, clearly demonstrating the large number of extreme cases.

Conduct Problems subscale

Conduct Problems item thresholds were ordered, items were locally independent, person fit and unidimensionality were acceptable. However, overall fit to the model was unsatisfactory (RMSEA >0.02, Table 2: analysis 3). Internal consistency was poor (PSI 0.10, α 0.65) with the subscale being able to discriminate between three strata.

Item fit residuals were within acceptable range though two had significant chi-squares (items 5 and 18).

Statistically significant DIF by ethnicity was present for item 12 and by gender for item 7. These two items were split by ethnicity and gender respectively (Table 2: analysis 4), resulting in satisfactory fit residuals, one item with a significant chi-square, significant improvement in RMSEA (0.03) but poor internal consistency (PSI=0.11, splitting items leads to missing data and α cannot be calculated). The easiest item to endorse was item 5 and the hardest item 12 (Table 3). The split item locations show that for children with the same level of conduct problems item 12 is more readily endorsed when they are Pasifika and less readily endorsed when they are NZE (difference of 1.22 logits). Item 7 is endorsed more readily by parents of boys than girls (difference of 0.32 logits). Targeting showed a floor effect (Figure 2).

Hyperactivity subscale

Ordered thresholds, local independence, person fit and unidimensionality were observed for the Hyperactivity subscale; however, overall fit to the model and internal consistency was unsatisfactory (RMSE >0.02; PSI 0.30, α 0.48; subscale discriminates between 3 strata, Table 2: analysis 5). Item fit residuals were out of range for item 21 and item 25 had a significant chi-square. Uniform DIF was statistically significant by ethnicity in two items (15 and 21). These items were therefore split by ethnicity. This improved fit to the Rasch model (Table 2: analysis 6) and displayed better fit than when these two items were combined into a testlet. Item fit residuals were within acceptable range of -2.5/+2.5; only one item had a significant item chi-square statistic (Table 3), and RMSEA was close to 0.02. However, internal consistency remained poor (PSI=0.31). The easiest item to endorse was item 15 (for Asian children) and the hardest item 10. The split item locations show that, for children with the same level of hyperactivity problems, item 15 is more readily endorsed when they are Asian and less readily endorsed when they are NZE (difference of 0.52 logits). Item 21 is endorsed more readily by parents of NZE children than Pasifika children (difference of 0.47 logits, table 3). The targeting map showed a floor effect (Figure 2).

Peer problems subscale

Ordered thresholds, local independence, person fit and unidimensionality were observed. However, overall fit to the Rasch model and internal consistency were unsatisfactory (RMSEA >0.02; PSI negative value, α 0.51, the subscale is able to discriminate between two strata, Table 2: analysis 7). Item fit residuals were acceptable, although two items had significant chi-squares. One item (23) displayed uniform DIF by ethnicity. After splitting this item by ethnicity, fit improved; all item fit residuals were within range (item 14 chi-square was borderline statistically significant), RMSEA was close to 0.02. PSI values remained negative, however (Table 2: analysis 8). The easiest item was item 23 (for Asian children) and the hardest item 14. Item 23 was easier for Asian children and hardest for NZE children (difference of 1.10 logits, Table 3). Targeting showed a significant floor effect (Figure 2).

Prosocial subscale

The subscale met the requirements for threshold ordering, local independence, person fit and unidimensionality. Overall fit to the Rasch model and internal consistency were unsatisfactory

(RMSEA >0.02; PSI negative values, α 0.29, subscale able to discriminate between two strata, Table 2: analysis 9). Item fit residuals were within the -2.5/+2.5 range, though two had significant item chi-square statistics. There was no DIF. Item 17 was the easiest to endorse; item 4 was the hardest to endorse. A ceiling effect was observed in the person-item-threshold map (Figure 2).

Difficulty scale

Two items had disordered thresholds, however, this was not statistically significant and item response categories did not need to be combined. Some local dependency was present in two item pairs. Unidimensionality was observed (Table 2: analysis 10). Five item fit residuals were out of the acceptable range of -2.5/+2.5 and four items showed uniform DIF by ethnicity (items 12, 16, 21 and 23). To examine whether DIF was present at the test level these items were combined into a testlet. This resulted in an absence of DIF, however, one item pair remained locally dependent (items 2 and 10). A second testlet was created to deal with this local dependency. The resulting scale was unidimensional, with locally independent items (Table 2: analysis 11). The RMSEA was within range suggesting overall fit to the Rasch model. Internal consistency was good (PSI 0.71, α 0.77, the scale was able to discriminate between six distinct strata). The fit residual for one item was slightly out of range (item 15, -2.777), however, given the negative value of this residual this indicates redundancy rather than misfit and the item was therefore retained. The easiest item to endorse was item 15, the hardest item 14. The person-item threshold map showed a normal distribution, although located to the left of the item locations on the latent trait. A conversion table was produced, which can be used to convert the raw ordinal score to an interval scale (Table 4).

Discussion

This study has shown that the SDQ items response categories work well, however, the five subscales diverge significantly from the Rasch model and four SDQ subscales include items that are biased by key variables with ethnicity having the greatest contribution. This raises critical questions about cultural equivalence. The five subscales suffer from a floor and ceiling effect and their internal consistency statistics are well below the acceptable range. By contrast, the total Difficulty scale, which combines the four subscales capturing children's problems, is unidimensional, fits the Rasch model (after dealing with DIF and local dependency) and has internal consistency sufficient to distinguish between six groups of children. The study has also shown that parents and teachers score children in their care differently. Thus, all three study hypotheses are rejected. This section will discuss our findings in terms of fit to the Rasch model, internal consistency, cultural equivalence and cross-informant reliability.

Fit to the Rasch model

The total Difficulty scale did fit the Rach model, after dealing with four DIF items and two locally dependent items. This scale has good internal consistency and is able to discriminate between six groups of children on the latent trait. We observed the population distribution, whilst following a normal pattern, was to the left of the item locations on the latent trait. Thus, the precision of person estimates at the lower of the scale will not be as good as for those at the higher end of the scale. However, the SDQ is used for screening and arguably precise measurement at the lower end is not needed, since all one needs to establish is that the child does not need to be referred for further assessment or intervention. As we achieved fit to the Rasch model we were able to provide a conversion table which can be used by clinicians to convert the raw ordinal score to more accurate interval level and which takes account of DIF.

Internal consistency

The 5 subscales are relatively short, which affects internal consistency and the subscales' ability to make fine distinctions between groups of people on the underlying trait.[25] In addition, there was significant divergence between the PSI and Cronbach's alpha statistics, with PSI being much smaller than alpha. This divergence can be explained by the way these statistics are calculated. The calculation of Cronbach's alpha assumes all standard errors (SEs) for individuals are the same, making it not a very robust statistics for skewed data.[45] This assumption results in relatively high values even in the presence of extreme scores and the Cronbach alpha values are therefore meaningless for SDQ data. This issue has not been raised in the SDQ literature; indeed, Cronbach's alpha values are widely reported as satisfactory.[49] In Rasch analysis the SE for every individual is estimated and the calculation of the PSI statistic takes these into account. Since SEs are largest for people with extreme scores, PSI will be smaller than Cronbach's alpha as observed in our skewed data. However, the purpose of the SDQ is to identify those children who would benefit from further assessment or intervention. Thus, the fact that we observed a floor and ceiling effect is not necessarily problematic.

Cultural equivalence

This study examined invariance by ethnicity at the item level and found lack of cultural equivalence. DIF (especially by ethnicity) was found for all the four subscales measuring problems, suggesting there are a number of questions to which parents respond differently despite overall scoring the same amount of problems on the trait being measured. The only other Rasch analysis study we were able to locate (conducted on data from children aged 12 to 18) did not include a DIF analysis and thus we cannot compare our findings against theirs.[24] Lack of measurement invariance of the subscales has also been shown by others (albeit on older children than in our sample) when using a CFA approach.[50] [51] Richter et al. found varying factor loadings and thresholds between different ethnic Norwegians and minority ethnic groups of adolescents and concluded that the total difficulty score is a preferable.[50] Similarly, Ortuño-Sierra et al. demonstrated that measurement variance was only partial, with 11 of the 25 items not being variant across different European samples.[51] By contrast, others have shown measurement invariance between British Indian and British white children using multi-group confirmatory factor analyses and demonstrated evidence of acceptable fit across ethnicity, although again their population was older (5 to 16 years) than the sample considered here.[52]

If measurement variance (DIF) is ignored, the child's difficulties can be over- or underestimated since the difficulty of the item varies by ethnic group, potentially leading to inaccurate identification of cases. This is important, given caseness has been shown to vary for different ethnic groups within the same country, and between countries.[53-55] Our study is unable to assess why such DIF occurs, since the study drew on secondary data. However, we can pose some possible factors that may have affected measurement variance, as discussed below.

Our recent qualitative study suggests there is variation in the way the SDQ is administered – some parents complete the tool by themselves and others receive support from nurses, possibly impacting on the way questions are interpreted.[29] In addition, New Zealand pre-school parents from Māori, Pacific Island, Asian, and new immigrant groups questioned the cultural validity of the SDQ.[29] Respondents in an Australian qualitative study exploring the SDQ in Aboriginal community-controlled health services reported that the use of a questionnaire as opposed to a general conversation or interview was deemed culturally inappropriate and that inter-relationships with peers were considered of less importance than relationships with family and participants.[56]

There are 85 different language versions available from the Youth in Mind website, though not one in Te Reo Māori (http://www.sdqinfo.org/). Translations and adaptations are not permitted without the involvement of that study team, which provides confidence in the robustness of translations. However, for our study we do not know whether respondents were offered the SDQ in the language of their choice, as such data are not collected as part of the B4SC. The literature includes six studies that examined and demonstrated some issues with SDQ translations.[13] Using a language version that is not understood by respondents will affect validity,[57] which may have occurred here.

It is possible that poor literacy impacts on answering the SDQ, as found by others.[58 59] In New Zealand there are many people (in proportion) with poorer than average literacy skills.[60] In addition, 18.6% of the New Zealand population report speaking two or more languages, the majority being born overseas (60.4%); many among these will have English as a second language.[61]

These aspects have particular relevance for Māori whānau (extended families) in New Zealand where is it estimated that 20% of Māori children and youth have Conduct problems.[62] Therefore, it is important that screening of Māori children during the preschool years is accurate in ensuring that Māori whānau both receive the support they need and at the same time are not pathologised by false positive findings. The 2013 New Zealand Census found that 21% of the almost 700,000 Māori population could hold conversation about everyday things in Te Reo Māori, which has been a national official language since 1987.[63] Yet, there is not Māori version of the SDQ, or a New Zealand version incorporating commonly used Māori words.

Cross-informant reliability

Cross-informant reliability was examined with ICCs which were well below the acceptable cut-off value of 0.6 (the mean in our study was 0.126). However, some argue that correlation coefficients can be used in the assessment of cross-informant reliability of the SDQ since parents and teachers make SDQ ratings based on different sources of information.[7 49] Our systematic literature review found weighted averages of coefficients between different informants ranged from 0.24 to 0.45,[13] similar to findings by others (range 0.26 to 0.47).[49] In our study the mean correlation coefficient was 0.28, meaning only 8% of the variance can be explained by scores from different informants. This implies the importance of taking into account the views of both parents and teachers when making a decision for onward referral, a practice that is not commonplace in New Zealand.[64]

A key strength of this study is the inclusion of all pre-school children in New Zealand for whom an SDQ assessment was available in 2011, resulting in our ability to assess the validity of the tool at the population level, with sufficient power to make sounds conclusions and ability to generalise to the wider New Zealand pre-school population. Another strength was robust data quality checks, and the exclusion of 39% of cases for which we had some concerns about quality (it being incomplete or containing multiple inconsistencies). From our steering group meetings we gathered that there were a few reasons underlying these quality issues. In some DHBs staff enter only the total scores, as opposed to item-level data. This practice leads to potential summing errors of total scores and these could not be checked, or indeed analysed (hence we excluded these cases). Secondly, some DHBs told us they set the default values of answers as zero rather than blank. Consequently, when there were missing data (for example if a teacher-completed SDQ was not available), the software would have summed these and arrived at total scores of zero. Given that the Prosocial scale is scored in the opposite direction of the others, zero scores on all subscales would be highly inconsistent and therefore shed doubt on data quality (and hence these were also excluded). An additional limitation was our inability to assess DIF by other key variables that may affect validity, e.g. first language or country of birth, as such data were not available.

In conclusion, the total Difficulty scale is internally valid and has acceptable internal consistency. Clinicians should use the conversion table as it accounts for bias by ethnic group. The five subscales are not valid and not suitable for use in their own right in New Zealand. Since consistency of scores between parents and teachers was poor it is advisable to use both parents and teachers' feedback when considering children's needs for referral to further assessment. Future work should examine whether validity is affected by different language versions used (in the same country).

Author contributions

PK conceived of the study, led on study design, project management, data analysis and dissemination. AV, HE, KMcP contributed to study design. AV contributed to the data analysis. PK drafted the manuscript and is the guarantor. All authors revised it critically for important intellectual content and approved the final version for publication. All authors agree to be accountable for all aspects of the work.

Competing Interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare: PK, AV, HE, KMcP had financial support from the Ministry of Health of New Zealand for the submitted work; subsequent to the completion of this project and data analysis KMcP became the Chief Executive of the Health Research Council of New Zealand; all other authors declare no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work. The views and opinions expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the funder.

Data sharing

Quantitative data from the study can be obtained from the author, subject to the funder's permission.

Figure 1. Item Characteristics Curves for items from the Strengths and Difficulties Questionnaire (parents, n=1,000)

Figure 2. Person-item-threshold maps Strengths and Difficulties Questionnaire (parents, n=1,000)

Box 1. Calculation of Root Mean Square Error of Approximation (RMSEA)

In Rasch analysis, RMSEA is calculated as follows:

RMSEA = $\sqrt{}$ ([((χ^2/df) - 1)/(N - 1)] , 0) [32]

χ^2 is the item-trait interaction chi-square (obtained from the analysis within the Rasch software),

df is its degrees of freedom

N is the sample size.

Notice that the RMSEA has an expected value of zero when the data fit the model. Overfit of the data to the model, χ^2/df < 1, is ignored. For a given χ^2, RMSEA decreases as sample size (N) increases.

Table 1. Intraclass correlation coefficients SDQ subscales, overall and by ethnicity (n=17,006)

Variable	**Ethnicity**				
	Overall*	**Māori**	**NZ European***	**Pasifika**	**Asian**
	r	**r**	**r**	**r**	**r**
Valid N	17056	2677	10735	1144	1169
Mean item correlations	0.282	0.237	0.315	0.130	0.210
Minimum item correlations	0.199	0.151	0.220	-0.009	0.055
Maximum item correlations	0.418	0.358	0.447	0.275	0.377
	ICC	**ICC**	**ICC**	**ICC**	**ICC**
Emotional Symptoms	0.126	0.067	0.186	0.017	0.098
Conduct Problems	0.137	0.112	0.179	0.038	0.079
Hyperactivity	0.174	0.136	0.245	0.050	0.122
Peer problems	0.139	0.100	0.202	0.004	0.162
Prosocial	0.055	0.048	0.066	0.040	0.035
Mean ICC	0.126	0.093	0.175	0.030	0.099
Minimum ICC	0.055	0.048	0.066	0.004	0.035
Maximum ICC	0.174	0.136	0.245	0.050	0.162

Table 2. Fit to the Rasch model – Strengths and Difficulties Questionnaire parents (SDQ-P) (n=1,000)

Subscales Analysis name			**Item Fit Residual**		**Person Fit** **Residual**		**Chi Square Interaction**			**RMSEA**[%]	**Internal consistency**[§]		**Unidimensionality** **T-Tests (CI)**[$$]
		N	**Mean**[$]	**SD**	**Mean**	**SD**	**Value**	**df**	**P**		**PSI** **Without** **extremes**	**α** **Without** **extremes**	**% (95% CI)**
Emotional Symptoms													
1	Initial	1,000	-0.791	0.894	-0.327	0.783	83.6	20	<0.0001	0.068	-0.40	0.15	0
2	Split items 16&24	1,000	-0.545	0.841	-0.343	0.735	99.1	41	<0.0001	0.045	-0.41	-	0
Conduct Problems													
3	Initial	1,000	0.266	1.273	-0.253	0.876	71.6	20	<0.0001	0.060	0.10	0.65	0
4	Split items 7&12	1,000	0.134	0.902	-0.254	0.882	75.3	45	0.003	0.031	0.11	-	0
Hyper-activity													
5	Initial	1,000	0.260	2.348	-0.359	1.147	97.3	25	<0.0001	0.06	0.30	0.48	0.5 (-1.0 to 2.0)
6	Split items 15&21	1,000	0.323	1.480	-0.365	1.134	125.6	69	<0.0001	0.03	0.31	-	0.5 (-1.0 to 2.0)
Peer Problems													

7	Initial	1,000	-0.339	0.868	-0.207	0.719	69.0	20	<0.0001	0.06	-0.49	0.51	0
8	Split item 23	1,000	-0.207	0.652	-0.213	0.733	79.5	52	0.008	0.03	-0.43	-	0
Prosocial													
9	Initial	1,000	-0.075	1.592	-0.319	1.079	66.6	20	<0.0001	0.06	-0.03	0.29	0.1 (-1.5 to 1.8)
Difficulty													
10	Initial	1,000	-0.448	1.848	-0.248	1.004	296.3	180	0.0001	0.03	0.71	0.79	5.9 (4.6 to 7.3)
11	Testlets DIF[%%] items & LD[§§] items	1,000	-0.615	1.321	-0.294	0.985	200.4	144	0.001	0.02	0.71	0.77	3.0 (1.6 to 4.4)

***Note* - Indices indicative of fit:**

[$] Mean item and person fit residuals: should be close to 0 (and <0.4); SD close to 1 (and <1.4)

[%] RMSEA (Root Mean Square Error of Approximation) <0.02

[§] Internal consistency PSI and α ≥0.70 (allows for group comparisons) and ≥0.85 (allows for individual clinical use)

[$$] Unidimensionality indicated if fewer than 5% of t-tests are significant (i.e. the 95% CI should include 5%)

[%%] DIF: Differential item Functioning

[§§] LD: Local Dependency

Table 3. Item locations (in location order) and fit statistics Strengths and Difficulties Questionnaire parents (SDQ-P) subscales (n=1,000)

Subscale & Items	**Location**	**SE**	**Fit**	**Chi Square**	**df**	**P**
			Residual	**value**		
Emotional problems [$]						
16 Māori	-0.871	0.113	-0.226	2.968	4	0.5631
16 NZE	-0.692	0.124	-0.036	0.60	3	0.8960
16 Asian	-0.538	0.118	-0.101	0.77	3	0.8569
16 Pasifika	-0.450	0.120	0.911	0.61	3	0.8936
24 Asian	-0.250	0.124	-0.185	5.13	3	0.1629
24 Māori	0.010	0.117	-0.737	9.69	4	0.0461
24 Pasifika	0.024	0.124	-0.002	11.857	3	0.0079
24 NZE	0.243	0.127	-1.610	14.095	3	0.0028
3	0.653	0.070	-0.615	15.156	5	0.0097
8	0.908	0.075	-1.970	21.479	5	0.0007
13	0.965	0.080	-1.423	16.749	5	0.0050
Conduct Problems [%]						
5	-0.985	0.063	0.011	15.38	5	0.0089
18	-0.707	0.066	-0.352	22.19	5	0.0005
7 Male	-0.594	0.096	1.209	7.71	5	0.1732
7 Fem	-0.271	0.100	1.917	6.09	5	0.2975
22	-0.012	0.072	0.156	8.49	5	0.1312
12 Pasifika	0.089	0.143	-0.148	3.527	5	0.6193
12 Māori	0.339	0.145	-0.512	5.862	5	0.3199
12 Asian	0.838	0.202	-0.030	2.344	5	0.7998
12 NZE	1.304	0.211	-1.049	3.733	5	0.5884
Hyperactivity [$]						
15Asian	-0.491	0.109	-0.395	8.25	5	0.1432
15 Māori	-0.315	0.117	0.433	1.78	6	0.9388

21 NZE	-0.234	0.142	2.204	17.50	5	0.0037
2	-0.206	0.056	-1.327	23.29	9	0.0056
21 Asian	-0.186	0.124	1.414	8.216	5	0.1447
15 Pasifika	-0.019	0.121	0.388	8.775	5	0.1184
15 NZE	0.032	0.126	-1.737	12.772	5	0.0256
21 Māori	0.114	0.129	1.743	7.403	6	0.2852
21 Pasifika	0.234	0.122	1.393	5.986	5	0.3076
25	0.360	0.066	1.421	9.335	9	0.4070
10	0.712	0.065	-1.984	22.26	9	0.0081
Peer Problems [%]						
23 A	-0.968	0.109	-0.571	1.959	4	0.7432
23 P	-0.870	0.107	0.307	4.311	5	0.5056
23 M	-0.217	0.119	0.038	5.529	4	0.2372
6	-0.026	0.065	0.526	10.572	9	0.3062
23 N	0.130	0.154	0.093	3.548	3	0.3147
11	0.233	0.066	-1.419	17.787	9	0.0377
19	0.491	0.071	0.131	12.305	9	0.1967
14	1.227	0.084	-0.763	23.501	9	0.0052
Prosocial [§]						
1	-0.487	0.079	-1.530	18.205	4	0.0011
4	-0.036	0.073	-0.273	12.624	4	0.0133
9	0.000	0.072	1.092	6.74	4	0.1502
17	0.008	0.071	-1.633	21.52	4	0.0003
20	0.515	0.073	1.972	7.52	4	0.1109
Difficulty [$$]						
15	-0.835	0.054	-2.777	27.39	9	0.0012
LD items [%%]	-0.606	0.037	-1.744	14.01	9	0.1221
5	-0.583	0.056	-0.595	8.71	9	0.4645
DIF items [§§]	-0.375	0.031	-2.500	21.03	9	0.0125
25	-0.331	0.061	0.036	14.05	9	0.1207

24	-0.314	0.058	0.839	7.44	9	0.5911
18	-0.313	0.059	-0.742	6.83	9	0.6553
6	-0.137	0.061	1.137	4.47	9	0.8777
7	-0.026	0.063	-1.305	23.26	9	0.0057
11	0.117	0.067	0.862	9.76	9	0.3702
22	0.308	0.068	-1.218	14.07	9	0.1199
3	0.311	0.071	1.017	11.50	9	0.2433
19	0.413	0.072	-1.247	10.59	9	0.3048
8	0.561	0.077	0.105	4.79	9	0.8525
13	0.646	0.087	0.621	9.37	9	0.4035
14	1.164	0.084	-2.326	13.15	9	0.1560

Note

[$] Bonferroni corrections applied P is statistically significant if <0.005

[%] Bonferroni corrections applied P is statistically significant if <0.006

[§] Bonferroni corrections applied P is statistically significant if <0.01

[$$] Bonferroni corrections applied P is statistically significant if <0.003

[%%] LD (Locally Dependent) items; combined into a testlet (item 2 and 10)

[§§] DIF (Differential Item Functioning) items combined into a testlet (items 12, 16, 21, 23)

Table 4. Conversion table for the Difficulty scale of the Strengths and Difficulties Questionnaire parents (SDQ-P)

Original Total Difficulty score (ordinal data)	Logit scores (interval level data)	Converted logit scores to 0-40 scale (interval level data)
0	-4.483	0
1	-3.655	4
2	-3.082	7
3	-2.685	8
4	-2.375	10
5	-2.117	11
6	-1.895	12
7	-1.699	13
8	-1.522	14
9	-1.36	15
10	-1.209	15
11	-1.068	16
12	-0.935	16
13	-0.809	17
14	-0.687	18
15	-0.571	18
16	-0.457	19
17	-0.347	19
18	-0.24	20
19	-0.134	20
20	-0.029	21
21	0.075	21
22	0.178	22
23	0.282	22

24	0.386	23
25	0.492	23
26	0.599	24
27	0.709	24
28	0.822	25
29	0.94	25
30	1.064	26
31	1.196	26
32	1.337	27
33	1.491	28
34	1.663	29
35	1.859	29
36	2.09	31
37	2.373	32
38	2.746	34
39	3.301	36
40	4.125	40

References

1. Eivers AR, Brendgen M, Borge AIH. Stability and change in prosocial and antisocial behavior across the transition to school: Teacher and peer perspectives. Early Education and Development 2010;**21**(6):843-64.

2. White J, Connelly G, Thompson L, et al. Assessing wellbeing at school entry using the Strengths and Difficulties Questionnaire: Professional perspectives. Educational Research 2013;**55**(1):87-98.

3. Kim-Cohen J, Caspi A, Moffitt TE, et al. Prior juvenile diagnoses in adults with mental disorder: Developmental follow-back of a prospective-longitudinal cohort. Archives of General Psychiatry 2003;**60**(7):709-17.

4. Kim-Cohen J, Arseneault L, Newcombe R, et al. Five-year predictive validity of DSM-IV conduct disorder research diagnosis in 41/2-5-year-old children. European Child and Adolescent Psychiatry 2009;**18**(5):284-91.

5. Bierman KL, Coie J, Dodge K, et al. School Outcomes of Aggressive-Disruptive Children: Prediction From Kindergarten Risk Factors and Impact of the Fast Track Prevention Program. Aggressive Behavior 2013;**39**(2):114-30.

6. Doughty C. The effectiveness of mental health promotion, prevention and early intervention in children, adolescents and adults. NZHTA Report ; 8(2). 2005.

7. Goodman R. The strengths and difficulties questionnaire: A research note. Journal of Child Psychology and Psychiatry and Allied Disciplines 1997;**38**(5):581-86.

8. Goodman R, Meltzer H, Bailey V. The Strengths and Difficulties Questionnaire: a pilot study on the validity of the self-report version. European Child & Adolescent Psychiatry 1998;**7**(3):125-30.

9. Ministry of Health. The B4 School Check. A handbook for practitioners. Wellington, 2008.

10. Williamson A, Este C, Clapham K, et al. What are the factors associated with good mental health among Aboriginal children in urban New South Wales, Australia? Phase I findings from the

Study of Environment on Aboriginal Resilience and Child Health (SEARCH). BMJ Open 2016;**6**(7).

11. Embretson SE, Reise SP. *Item Response Theory for Psychologists*. London: Lawrence Erlbaum Associates, Publishers, 2000.
12. Cano S, Klassen AF, Scott A, et al. Health outcome and economic measurement in breast cancer surgery: Challenges and opportunities. Expert Review of Pharmacoeconomics and Outcomes Research 2010;**10**(5):583-94.
13. Kersten P, Czuba K, McPherson KM, et al. A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. International Journal of Behavioral Development 2016;**40**(1):64-75.
14. Andrich D. A rating formulation for ordered response categories. Psychometrika 1978;**43**:561-73.
15. Rasch G. *Probabilistic models for some intelligence and attainment tests (revised and expanded ed.)*. Chicago: The University of Chicago Press, 1960/1980.
16. Bond TG, Fox CM. *Applying the Rasch model. Fundamental measurement in the human sciences*. London: Lawrence Erlbaum Associates, 2001.
17. Klein AM, Otto Y, Fuchs S, et al. Psychometric properties of the parent-rated SDQ in preschoolers. European Journal of Psychological Assessment 2013;**29**(2):96-104.
18. Tobia V, Gabriele MA, Marzocchi GM. The Italian Version of the Strengths and Difficulties Questionnaire (SDQ)-Teacher: Psychometric Properties. Journal of Psychoeducational Assessment 2013;**31**(5):493-505.
19. Mieloo CL, Bevaart F, Donker MC, et al. Validation of the SDQ in a multi-ethnic population of young children. The European Journal of Public Health 2014;**24**(1):26-32.
20. Borg A-M, Pälvi K, Raili S, et al. Reliability of the Strengths and Difficulties Questionnaire among Finnish 4-9-year-old children. Nordic Journal of Psychiatry 2012;**66**(6):403-13.
21. Goodman A, Lamping DL, Ploubidis GB. When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the strengths and difficulties questionnaire (SDQ): Data from british parents, teachers and children. Journal Of Abnormal Child Psychology 2010;**38**(8):1179-91.

22. Wright BD. Comparing Rasch measurement and factor analysis. Structural Equation Modeling 1996;**3**:3-24.

23. Christensen KB, Engelhard J, G., Salzberger T. Rasch vs. Factor Analysis. Rasch Measurement Transactions 2012;**26**(3):1373-8.

24. Hagquist C. The psychometric properties of the self-reported SDQ **-** An analysis of Swedish data based on the Rasch model. Personality and Individual Differences 2007;**43**(5):1289-301.

25. Streiner DL, Norman GR. *Health Measurement Scales: a practical guide to their development and use*. Oxford: Oxford University Press, 2008.

26. Statistics New Zealand. Māori Population Estimates: Mean year ended 31 December 2016. Secondary Māori Population Estimates: Mean year ended 31 December 2016 2016. http://www.stats.govt.nz/browse_for_stats/population/estimates_and_projections/MaoriPopulationEstimates_HOTPMYe31Dec16.aspx.

27. Høegh MC, Høegh S-M. Trans-adapting outcome measures in rehabilitation: Cross-cultural issues. Neuropsychological Rehabilitation 2009;**19**(6):955-70.

28. de Klerk G. Cross-Cultural Testing In: Born M, Foxcroft, C.D., Butter, R., ed. Online Readings in Testing and Assessment: International Test Commission, 2008.

29. Kersten P, Dudley M, Nayar S, et al. Cross-cultural acceptability and utility of the strengths and difficulties questionnaire: views of families. BMC Psychiatry 2016;**16**(1):347.

30. Andrich D. *Rasch models for measurement series: quantitative applications in the social sciences no. 68*. London: Sage Publications, 1988.

31. Linacre JM. Rasch Power Analysis: Size vs. Significance: Infit and Outfit Mean-Square and Standardized Chi-Square Fit Statistic. Rasch Meas Trans 2003;**17**(1):918.

32. Tennant A, Pallant JF. The Root Mean Square Error of Approximation (RMSEA) as a supplementary statistic to determine fit to the Rasch model with large sample sizes. Rasch Meas Trans 2012;**25**(4):1348-9.

33. Linacre JM. Sample size and item calibration [or Person Measure] stability. Rasch Measurement Transactions 1994;**7**(4):328.

34. Wright BD, Tennant A. Sample size again. Rasch Measurement Transactions 1996;**9**(4):468.

35. Charter RA, Feldt LS. Confidence intervals for true scores: Is there a correct approach? Journal of Psychoeducational Assessment 2001;**19**(4):350-64.

36. Nunnally JC, Bernstein IH. *Psychometric theory (3rd ed.)*. New York: McGraw-Hill, 1994.

37. RUMM2030 [program]: RUMM Laboratory Pty Ltd., 2009.

38. Marais I, Andrich D. Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. Journal of Applied Measurement 2008;**9**(2):105-24.

39. Wainer H, Kiely G. Item clusters and computer adaptive testing: A case for testlets. J Educ measurement 1987;**24**:185-202.

40. Grimby G. Useful reporting of DIF. Rasch Measurement Transactions 1998;**12**(3):651.

41. Holland PW, Wainer H. Differential Item Functioning. NJ: Hillsdale. Lawrence Erlbaum, 1993.

42. Smith EV. Detecting and evaluation the impact of multidimensionality using item fit statistics and principal component analysis of residuals. Journal of Applied Measurement 2002;**3**:205-31.

43. Tennant A, Pallant JF. Unidimensionality matters! (a tale of two Smiths?). Rasch Measurement Transactions 2006;**20**:1048-51.

44. Wright BD. Reliability and separation. Rasch Measurement Transactions 1996;**9**(4):472.

45. Sheng Y, Sheng Z. Is Coefficient Alpha Robust to Non-Normal Data? Frontiers in Psychology 2012;**34**(1):1-13.

46. Wright BD. Separation, Reliability and Skewed Distributions: Statistically Different Levels of Performance. Rasch Meas Trans 2001;**14**(4):786.

47. Andrich D. Rating formulation for ordered response categories. Psychometrica 1978;**43**(4):561-73.

48. Masters G. A Rasch model for partial credit scoring. Psychometrica 1982;**47**:149-74.

49. Stone LL, Otten R, Engels RC, et al. Psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire for 4- to 12-year-olds: a review. Clinical Child And Family Psychology Review 2010;**13**(3):254-74.

50. Richter J, Sagatun A, Heyerdahl S, et al. The Strengths and Difficulties Questionnaire (SDQ) - Self-Report. An analysis of its structure in a multiethnic urban adolescent sample. Journal of Child Psychology and Psychiatry and Allied Disciplines 2011;**52**(9):1002-11.

51. Ortuño-Sierra J, Fonseca-Pedrero E, Aritio-Solana R, et al. New evidence of factor structure and measurement invariance of the SDQ across five European nations. European Child and Adolescent Psychiatry 2015;**24**(12):1523-34.

52. Goodman A, Patel V, Leon DA. Why do British Indian children have an apparent mental health advantage? Journal of Child Psychology and Psychiatry and Allied Disciplines 2010;**51**(10):1171-83.

53. Goodman A, Heiervang E, Fleitlich-Bilyk B, et al. Cross-national differences in questionnaires do not necessarily reflect comparable differences in disorder prevalence. Social Psychiatry And Psychiatric Epidemiology 2012;**47**(8):1321-31.

54. de Vries PJ, Davids EL, Mathews C, et al. Measuring adolescent mental health around the globe: psychometric properties of the self-report Strengths and Difficulties Questionnaire in South Africa, and comparison with UK, Australian and Chinese data. Epidemiology and Psychiatric Sciences 2017:1-12.

55. Kersten P, Vandal AC, Elder H, et al. Concurrent Validity of the Strengths and Difficulties Questionnaire in an Indigenous Pre-School Population. Journal of Child and Family Studies 2017;**26**(8):2126-35.

56. Williamson A, Redman S, Dadds M, et al. Acceptability of an emotional and behavioural screening tool for children in Aboriginal Community Controlled Health Services in urban NSW. Australian and New Zealand Journal of Psychiatry 2010;**44**(10):894-900.

57. Beaton DE, Bombardier C, Guillemin F, et al. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine 2000;**25**(24):3186-91.

58. Samad L, Hollis C, Prince M, et al. Child and adolescent psychopathology in a developing country: Testing the validity of the Strengths and Difficulties Questionnaire (Urdu version). International Journal Of Methods In Psychiatric Research 2005;**14**(3):158-66.

59. Thabet AA, Stretch D, Vostanis P. Child mental health problems in Arab children: Application of the strengths and difficulties questionnaire. International Journal of Social Psychiatry 2000;**46**(4):266-80.

60. Lane C. Adult literacy and numeracy in New Zealand – A regional analysis. Perspectives from the Adult Literacy and Life Skills Survey, 2012.

61. Statistics New Zealand. 2013 Census QuickStats about culture and identity. Available from http://www.stats.govt.nz. Wellington, New Zealand, 2014.

62. The Advisory Group on Conduct Problems. Conduct problems. Best practice report. Wellington, 2009.

63. Statistics New Zealand. Te Kupenga, 2013.

64. Hedley C, Thompson S, Morris Mathews K, et al. The B4 School Check behaviour measures: Findings from the hawke's bay evaluation. Nursing Praxis in New Zealand 2012;**28**(3):13-23.

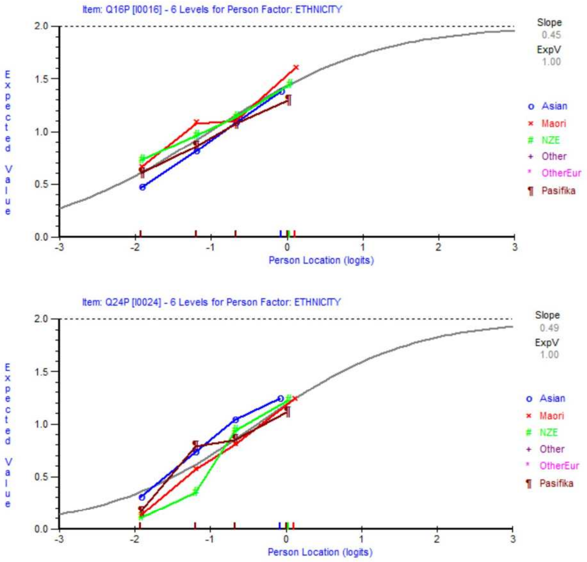
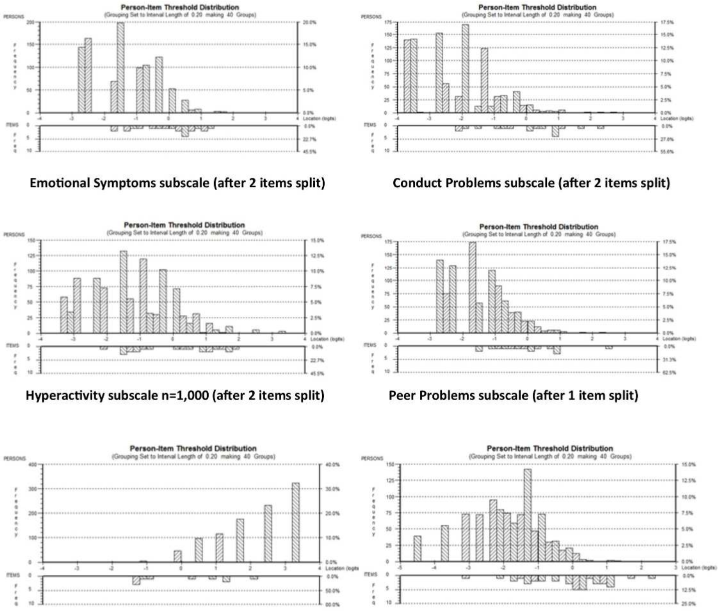


Figure 1. Item Characteristics Curves for items from the Strengths and Difficulties Questionnaire (parents, n=1,000)

209x297mm (300 x 300 DPI)



Emotional Symptoms subscale (after 2 items split)

Conduct Problems subscale (after 2 items split)

Hyperactivity subscale n=1,000 (after 2 items split)

Peer Problems subscale (after 1 item split)

Prosocial subscale (no items split)

Difficulty scale

Figure 2. Person-item-threshold maps Strengths and Difficulties Questionnaire (parents, n=1,000)

209x297mm (300 x 300 DPI)

STROBE Statement—checklist of items that should be included in reports of observational studies

	Item No	Recommendation	Page number
Title and abstract	1	(*a*) Indicate the study's design with a commonly used term in the title or the abstract	2
		(*b*) Provide in the abstract an informative and balanced summary of what was done and what was found	2
Introduction			
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported	4-6
Objectives	3	State specific objectives, including any prespecified hypotheses	6
Methods			
Study design	4	Present key elements of study design early in the paper	2, 6
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection	6-8
Participants	6	(*a*) *Cohort study*—Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up *Case-control study*—Give the eligibility criteria, and the sources and methods of case ascertainment and control selection. Give the rationale for the choice of cases and controls *Cross-sectional study*—Give the eligibility criteria, and the sources and methods of selection of participants	N/A 7
		(*b*) *Cohort study*—For matched studies, give matching criteria and number of exposed and unexposed *Case-control study*—For matched studies, give matching criteria and the number of controls per case	N/A
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable	N/A

Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group	8
Bias	9	Describe any efforts to address potential sources of bias	7
Study size	10	Explain how the study size was arrived at	8
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why	7-8
Statistical methods	12	(*a*) Describe all statistical methods, including those used to control for confounding	8-11
		(*b*) Describe any methods used to examine subgroups and interactions	8-11
		(*c*) Explain how missing data were addressed	7
		(*d*) *Cohort study*—If applicable, explain how loss to follow-up was addressed *Case-control study*—If applicable, explain how matching of cases and controls was addressed *Cross-sectional study*—If applicable, describe analytical methods taking account of sampling strategy	N/A
		(*e*) Describe any sensitivity analyses	N/A

Continued on next page

Results			
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analysed	11
		(b) Give reasons for non-participation at each stage	N/A
		(c) Consider use of a flow diagram	N/A
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders	11
		(b) Indicate number of participants with missing data for each variable of interest	7
		(c) *Cohort study*—Summarise follow-up time (eg, average and total amount)	N/A
Outcome data	15*	*Cohort study*—Report numbers of outcome events or summary measures over time	N/A
		Case-control study—Report numbers in each exposure category, or summary measures of exposure	N/A
		Cross-sectional study—Report numbers of outcome events or summary measures	N/A
Main results	16	(*a*) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included	N/A
		(*b*) Report category boundaries when continuous variables were categorized	N/A
		(*c*) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period	N/A
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses	N/A
Discussion			
Key results	18	Summarise key results with reference to study objectives	15
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias	19
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from	16-19

		similar studies, and other relevant evidence	
Generalisability	21	Discuss the generalisability (external validity) of the study results	19
Other information			
Funding	22	Give the source of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based	3

*Give information separately for cases and controls in case-control studies and, if applicable, for exposed and unexposed groups in cohort and cross-sectional studies.

Note: An Explanation and Elaboration article discusses each checklist item and gives methodological background and published examples of transparent reporting. The STROBE checklist is best used in conjunction with this article (freely available on the Web sites of PLoS Medicine at http://www.plosmedicine.org/, Annals of Internal Medicine at http://www.annals.org/, and Epidemiology at http://www.epidem.com/). Information on the STROBE Initiative is available at www.strobe-statement.org.