

Age appropriateness of the self-report Strengths and Difficulties Questionnaire

Louise Black, Rosie Mansfield and Margarita Panayiotou

The University of Manchester

Black, L., Mansfield, R., & Panayiotou, M. (2020). Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire. *Assessment*.

<https://doi.org/10.1177/1073191120903382>

Author Note

Louise Black, Manchester Institute of Education, University of Manchester; Rosie Mansfield, Manchester Institute of Education, University of Manchester, Margarita Panayiotou, Manchester Institute of Education, University of Manchester.

**Acknowledgments:** The data used in this study were collected as part of the HeadStart learning program. The authors are therefore grateful for the work of the wider research teams at the Anna Freud Centre and the University of Manchester for their role in coordinating the evaluation, as well as collecting and managing the data. The authors also acknowledge the National Pupil Database from which demographic data were obtained. Finally, we are extremely grateful to all students who took part in this study, as well as the local authorities and schools for their help in recruiting them.

**Funding:** The data used in this study was collected as part of HeadStart learning program and supported by funding from the National Lottery Community Fund, grant R118420. The funders did not seek to influence any aspect of the secondary analysis reported in this study. The

content is solely the responsibility of the authors and it does not reflect the views of the National Lottery Community Fund. Louise Black is funded by the National Institute for Health Research and Rosie Mansfield is funded by the Department for Education.

Correspondence concerning this article should be addressed to Louise Black, Manchester Institute of Education, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK.  
Email: [louise.black@manchester.ac.uk](mailto:louise.black@manchester.ac.uk)

### Abstract

The self-report version of the strengths and difficulties questionnaire (SDQ) is widely used in clinical and research settings. However, the measure's suitability for younger adolescents has recently been called into question by readability analysis. To provide further insight into the age-appropriateness of the SDQ, readability was assessed at the item level alongside consideration of item quality criteria, its factor structure was analyzed, and measurement invariance between adolescents in year seven (age 11–12) versus year nine (age 13–15) was tested. The measure showed a wide range of reading ages, and the theorized factor structure was unacceptable. Measurement invariance was therefore considered for a flexible exploratory structural equation model, and no evidence of differences between age groups was found. Suggestions are made for the measure's revision based on these findings.

*Keywords:* strength and difficulties questionnaire; readability; measurement invariance; mental health; adolescents.

### Age appropriateness of the self-report Strengths and Difficulties Questionnaire

The self-report version of the Strengths and Difficulties Questionnaire (SDQ) is a popular measure of mental health in 11–16 year-olds (Goodman, Meltzer, & Bailey, 1998; Johnston & Gowers, 2005) that has been extensively used in epidemiological research (e.g., Hafekost et al., 2016; NHS Digital, 2018; Polanczyk, Salum, Sugaya, Caye, & Rohde, 2015). Self-report measures are generally attractive in research, particularly in longitudinal and large-scale studies. This is partly because young people can be easier to recruit than parents, and data burden is reduced compared to teacher report methods (Humphrey & Wigelsworth, 2016). Moreover, such measures allow direct assessment of the young person's perspective in accordance with policy recommendations (Deighton et al., 2014). Despite these advantages, scale- and subscale-level analysis suggest the SDQ may be unsuitable for those with reading ages below 13–14 (Patalay, Hayes, & Wolpert, 2018). Not only is this higher than the intended 11-year-old population, it also exceeds general scale development recommendations, which suggest that measures should never exceed the reading level of a 12-year-old (Terwee et al., 2007). There is also evidence to suggest that the reading age of individuals can be up to 5 grades lower than their reported education grade, especially for those experiencing mental health difficulties (Jackson et al., 1991; Jensen, Fabiano, Lopez-Williams, & Chacko, 2006). There is, therefore, a need for better understanding of the age appropriateness of this measure.

Though the self-report SDQ has been consistently employed in large national studies (e.g., Hafekost et al., 2016; NHS Digital, 2018), and has been recommended for research and clinical settings (Vostanis, 2006; Wolpert, Cheng, & Deighton, 2015), robust evidence of its factor structure is scant. Two review articles have broadly advocated for the use of the self-report

SDQ, as a well-validated measure (Vostanis, 2006; Wolpert et al., 2015). However, it should be noted that psychometric evidence underpinning their recommendations often related to translated versions, though psychometric characteristics are likely version dependent (Flake, Pek, & Hehman, 2017). Indeed, the self-report SDQ has shown only partial measurement invariance across different language versions (Ortuño-Sierra et al., 2015b). Furthermore, studies of the English version on which recommendations were made particularly failed to report model fit (Goodman, 2001; Goodman et al., 1998). Though exploratory factor analysis was used in the original study, a 5-factor solution was retained despite substantial cross-loadings for seven items (Goodman, 2001), suggesting potential problems with the structure. Where confirmatory factor analysis (CFA) techniques were employed to analyze the self-report English version, the proposed structure was also shown to be problematic, with inconsistent fit based on recommended guidelines. These suggest values of around .95 for the comparative fit index (CFI) and around .06 for the root mean square error of approximation (RMSEA) can be judged to be acceptable (Hu & Bentler, 1999). Goodman, Lamping, and Ploubidis (2010) found CFI = .837 and RMSEA = .063 via weighted least squares means and variance adjusted (WLSMV), while Percy, McCrystal, and Higgins (2008) reported CFI = .817 and RMSEA = .047 via robust maximum likelihood (MLR) estimation. The consistently low CFI may be due to problems with the pattern of covariances specified by the model, consistent with the known substantial cross-loadings (Goodman, 2001; Percy et al., 2008), though discrepancies between RMSEA and CFI can occur for many different reasons (see Lai & Green, 2016, for more details). The fact that both studies include adolescents as young as 11–12 may also have contributed to model misfit.

This lack of clear support for the self-report SDQ's factor structure suggests a need for more detailed examination of its psychometric qualities, as has been explicitly called for in a

recent systematic review (Bentley, Hartley, & Bucci, 2019). This is particularly necessary given the centrality of the measure in adolescent mental health research (e.g., Deighton et al., 2019; Dray et al., 2016; Hafekost et al., 2016; NHS Digital, 2018; Polanczyk et al., 2015; Wigelsworth, Humphrey, & Lendrum, 2012). Although evidence based on the SDQ suggests an increase in mental health difficulties in mid adolescence, around ages 14–15 (Deighton et al., 2019; Dray et al., 2016), it is not clear whether differences between early adolescents, around ages 11–12, and the 14–15 age group are due to differences in measurement properties, or the SDQ’s high reading age (Patalay et al., 2018). Indeed, measurement invariance between different age groups is yet to be examined, which we therefore sought to address in the current study. The choice of age groups in the current study was selected for pragmatic reasons since we conducted secondary data analysis. Nevertheless, the use of this dataset enabled examination of the key transition to mid adolescence. It also allowed comparison between the SDQ’s youngest intended age (11 years old), as per its original validation (Goodman et al., 1998), and the recommended minimum age (13 years old) based on recent readability findings (Patalay et al., 2018).

While the analysis of readability by Patalay et al. (2018) provided valuable insight into the age appropriateness of the measure, readability was only considered for whole subscales meaning three issues remain unexplored. First, while considering items together as subscales or whole measures allows the use of texts of more appropriate length for readability formulas, information is lost about individual items (Oakland & Lane, 2004). Second, the presentation of items in accordance with psychometric best practice, including factor structure, should also be considered. For instance, items should have appropriate response formats and consist of single statements to avoid confusion (Sarlis, 2014; Terwee et al., 2007). Finally, while age invariance of the proxy version has been considered (He, Burstein, Schmitz, & Merikangas, 2013),

measurement invariance of the self-report English instrument has not been tested, to our knowledge. Based on these identified gaps, we aimed to explore the following for the self-report SDQ: 1) item-level readability, 2) item quality, 3) the factor structure, and 4) age measurement invariance between English secondary school students in year seven (age 11–12) and year nine (age 13–15). We hypothesized the reading age to be higher than the intended population, consistent with Patalay et al. (2018) and that item quality would vary according to psychometric criteria (this has not been evaluated previously and was therefore exploratory). Given that findings on the structure of SDQ have been conflicting, we were unable to hypothesize which structure would be the most appropriate, thus the third aim of our study was also necessarily exploratory. Finally, we hypothesized non-measurement invariance between the two age groups, as we expected the year nine group to have a better understanding of the items, based on previous readability evidence (Patalay et al., 2018).

### **Method**

Secondary data analysis was conducted of a large project aimed at promoting resilience in six areas of England, chosen on the basis of need. The original dataset consisted of 30,842 students, though 552 cases were excluded (1.8%) from current analyses since these had missing data for all SDQ items. Students were in year seven (50.7%, aged 11–12,  $M = 12.21$ ,  $SD = 0.29$ ) and nine (49.3%, aged 13–15,  $M = 14.20$ ,  $SD = 0.29$ ) from 114 schools (52.4% female). The ethnicity of our sample was very similar to national figures (Department for Education, 2017b) with 74.1% white, 9.5% Asian, 5.7% Black, 3.9 Mixed, .2% Chinese, 1.5% any other ethnic background, and 1.2% unclassified. The proportion of pupils with a special educational need was 11.6%, compared to the national figure of 14.4% (Department for Education, 2017c). Rates of low income were above average in this community sample, given the focus of the project: The

percentage of students who had ever been eligible for free school meals was 36.4% which was above the national average of 29.1% for those eligible in the previous six years (Department for Education, 2017a).

Total difficulties scores for the SDQ were also above rates expected in community samples, based on the measure's 20 year-old bandings (Goodman et al., 1998): 62.2% scored in the 'normal' range compared to 80% in the validation sample, 18.4% scored in the 'borderline' range compared to 10% in the validation sample and 19.6% scored in the 'abnormal' range compared to 10% in the validation sample. However, self-reported psychological wellbeing in the current sample ( $M = 23.88$ ,  $SD = 5.33$ ) was similar to the average found in a nationally representative sample of 16-24 year-olds ( $M = 23.57$ ,  $SD = 3.61$ ; Ng Fat, Scholes, Boniface, Mindell, & Stewart-Brown, 2017). Reading ability was also below average based on end of primary school test results, with 63% of the year seven cohort reaching the expected grade compared to the national result of 66% (Department for Education, 2016), and 72.2% of the year nine cohort compared to the national result of 78% (Department for Education, 2014).

Following approval by the UCL Research Ethics Committee (UCL Ref: 8097/003) survey data were collected via a secure online portal during the normal school day from students whose parents had not opted-out. The SDQ was completed as part of a battery of measures, all of which had explanations for items found to raise issues during piloting. These were constructed to help pupils without altering items, and since researchers did not administer the survey face-to-face they could not respond to queries. Pupils were instructed that these could be obtained by hovering their mouse over certain words. For example, if pupils hovered over the word "restless", they were given the explanation "unable to stay still". All items which had explanations are indicated in Table 2.



Students responded to the 25-item SDQ (Goodman et al., 1998) using a 3-point Likert scale (*not true, somewhat true, certainly true*). These 25 items form five subscales of five items each (more detail on the content of items can be found in Table 2). Internal consistency coefficients are presented in several formats to reflect both the typically reported standard (Cronbach's alpha), as well as formulae that account for violations likely present in the data (see Table 1). Ordinal alpha accounts for the ordinal nature of Likert items since it is based on the polychoric correlation matrix (Gadermann, Guhn, & Zumbo, 2012), while McDonald's omega is a model-based reliability which does not assume tau-equivalence (Raykov & Marcoulides, 2016). In line with other assessments of the SDQ (Bøe, Hysing, Skogen, & Breivik, 2016; Panayiotou, Humphrey, & Wigelsworth, 2019), ordinal alpha and omega were shown to be higher than Cronbach's alpha in the current sample (see Table 1). ESEM factor loadings can also be found in supplemental table S1.

### Analysis

**Readability testing.** Calculating multiple readability estimates is recommended given the lack of a gold standard readability formula, and the variability in their focus (Janan & Wray, 2012). The current study applied four widely used and established readability assessments, all of which are calculated by incorporating different text components. The Dale-Chall Readability Formula (DC; Chall & Dale, 1995; Dale & Chall, 1948), considers the percentage of difficult words, and the average sentence length. Difficult words are those that do not appear on the Dale-Chall Readability word list:

$$DC = 0.1579(DW/TW \times 100) + 0.0496(AWS) + 3.6365$$

Where *DW* = total number of difficult words, *TW* = total number of words, *AWS* = average number of words per sentence.

The Flesch-Kincaid Reading Grade (FK; Kincaid, Fishburne, Rogers, & Chissom, 1975), considers average syllables per word and the average sentence length:

$$FK = (0.39 \times AWS) + (11.8 \times ASW) - 15.59$$

where *AWS* = average number of words per sentence; *ASW* = average number syllables per word.

The Gunning Fog Index (GFI; Gunning, 1952) considers number of words, sentences and hard words (those with three syllables or more):

$$GFI = 0.4 \times [ AWS + (100HW / TW) ]$$

where *AWS* = average number of words per sentence; *HW* = total number of hard words; *TW* = total number of words.

Finally, the Coleman Liau Index (CLI; Coleman & Liau, 1975) incorporates number of letters instead of syllables:

$$CLI = (0.0588 \times LW) - (0.296 \times SW) - 15.8$$

where *LW* = average number of letters per 100 words; *SW* = average number of sentences per 100 words.

All indices provide readability as a US grade-level. The readability of SDQ items and subscales was then calculated by averaging the US-grade level score of the four indices, and then adding six to get the average reading age. The age appropriateness of SDQ items was judged against the original minimum recommended age of 11 (Goodman et al., 1998).

**Item quality criteria.** Consistent with readability indices, psychometric guidance suggests scale items should be simple in language and grammar, regardless of the age of the target population (Irwing & Hughes, 2018; Terwee et al., 2007). Beyond this, other important aspects of the content and structure of items must be considered alongside readability tests, for a

more comprehensive assessment (Oakland & Lane, 2004). Additional item quality criteria deemed relevant to age-appropriateness and mental health were therefore identified to supplement readability analyses. First, items should ideally consist of single statements (Irwing & Hughes, 2018; Saris, 2014; Terwee et al., 2007), and avoid reverse wording to reduce confusion (Irwing & Hughes, 2018; van Sonderen, Sanderman, & Coyne, 2013). Floor and ceiling effects (endorsement of the lowest or highest response at > 15%) should not be present. Absence of these is an indication that measures reliably distinguish individuals across the range of symptoms (Terwee et al., 2007). Items should also be presented with a clear and appropriate reference period to the concept under study (Irwing & Hughes, 2018; Saris, 2014). Since all items had the same reference period, we used the first three criteria to assess items and considered those that satisfied two out of three to be of higher quality.

**Factor structure and measurement invariance.** Given the poor factor structure of the self-report SDQ in other samples (Goodman et al., 2010; Goodman, 2001; Percy et al., 2008), we considered both CFA and ESEM with geomin rotation (see Figure 1). We estimated three CFA models, the first of which was a correlated structure of the five subscales, based on the original theoretical structure of the measure representing the five subscales typically used (Goodman, 2001). Secondly, we included a correlated 2-factor higher-order structure in which emotional problems and peer problems loaded onto a second-order internalizing factor, and conduct problems and hyperactivity loaded onto a second-order externalizing factor as suggested elsewhere (Goodman et al., 2010). Thirdly, we estimated a bifactor model (Chen & Zhang, 2018) with a general difficulties factor, and four residual difficulty subdomain factors. This model has shown some promise in other language versions (e.g., Ortuño-Sierra, Chocarro, Fonseca-Pedrero, Riba, & Muñiz, 2015a) and allows the total difficulties subscale to be represented as a general

factor after accounting for specific variance captured by each of the four problem domains. The prosocial factor was excluded from both the bifactor and higher-order models since these were used to examine the hypothesized 4-factor total difficulties score (Goodman, 2001). We finally tested a 5-factor ESEM model, which was used to explore age measurement invariance.

Where measures lack proposed dimensionality, as is the case with the self-report SDQ (Goodman et al., 2010; Goodman, 2001; Percy et al., 2008), and invariance testing is warranted, given recent claims about age (Deighton et al., 2019; Dray et al., 2016; NHS Digital, 2018), exploratory structural equation modelling (ESEM) techniques can be used (Marsh, Nagengast, & Morin, 2013a). As others have pointed out, though ESEM structures should not be used to conceal problems with a measure, they can provide a more realistic framework for measurement invariance analysis where CFA models do not fit sufficiently well (Tóth-Király, Bõthe, Rigó, & Orosz, 2017). Furthermore, given the substantial cross-loadings and shared variance in the SDQ (Goodman, 2001; Percy et al., 2008), ESEM can provide a more robust approach than post-hoc addition of parameters (e.g., crossloadings) following modification indices (Chiorri, Hall, Casely-Hayford, & Malmberg, 2016). We therefore opted to extract five factors in line with the original theoretical model, but in ESEM every item is permitted to load onto every factor so that shared variance in the data is not misspecified.

When accounting for the fact data were sampled from pupils clustered in schools (using type = complex), the ESEM models required greater numbers of parameters to be estimated than there were schools in the sample ( $165 > 114$ ), thus resulting in a warning about the trustworthiness of standard errors. Given that the implications of this in model estimation are not well understood (Muthén & Muthén, 2016), and parameter estimates would not be directly affected, clustering effects were not controlled for. This decision was guided by the small intra-

cluster correlations for the SDQ variables ( $<.05$ ) and the fact that controlling for clustering made little difference to the standard errors and therefore conclusions (results can be provided upon request). For consistency we therefore did not account for clustering in any model.

Chi-square difference testing is typically used to compare the fit of measurement invariance models. However, its sensitivity to sample size made this inappropriate for our study, suggesting approximate fit indices should be used. Since the majority of measurement invariance simulations focusing on performance of fit indices have treated items as continuous (Chen, 2007; Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008), the degree to which common fit indices are appropriate for comparing models using polychoric matrices and WLSMV is unclear. For instance, given that the chi-square of WLSMV is not comparable in the same way as for maximum likelihood, CFI comparisons might not be appropriate in these cases (Sass, Schmitt, & Marsh, 2014). Analyses were therefore conducted in Mplus 8.3 using MLR and treating items as continuous. This also allowed us to account for the non-normality of the data, and enabled missing data to be handled via full information maximum likelihood under the assumption of missing at random (Muthén & Muthén, 1998-2017). All cases with data for at least one SDQ item were therefore included in our analysis. Though items were treated as continuous, floor effects were likely in a screening measure, so sensitivity tests for the CFA and ESEM models were conducted, in which items were treated as ordinal using WLSMV (Brown, 2015; Li, 2016).<sup>1</sup>

Model fit was judged in line with published recommendations. Chi-square statistics are reported, but not interpreted as indicating fit given their known sensitivity to sample size. The CFI and the Tucker Lewis index (TLI) were considered to be acceptable at around .95, and

---

<sup>1</sup> WLSMV solutions were not estimated for measurement invariance testing, given the problems with comparing CFI for this estimator (Sass et al., 2014).

RMSEA around .06 (Hu & Bentler, 1999). The standardized root mean squared residual (SRMR) was considered to be acceptable  $< .08$  in the absence of any large residuals (Asparouhov & Muthén, 2018). In addition to these standardized indices, the Akaike information criterion (AIC) and Bayesian information criterion (BIC) are also reported to compare models with the same outcome variables, with lower values indicating better model fit.

Measurement invariance testing was conducted by estimating baseline models for each age group separately, followed by a configural model in which parameters were freely estimated in each group, a metric model with loadings constrained to be equal across groups, and finally a scalar model in which intercepts were also held equal (Muthén & Muthén, 1998-2017). Given the large sample size, CFI difference ( $\Delta$ CFI) was used to judge approximate invariance (Sass et al., 2014). In line with wider ESEM literature (Marsh et al., 2013a; Marsh et al., 2013b; Tóth-Király et al., 2017), and specific invariance analysis of the SDQ (Chiorri et al., 2016), we adopted a threshold of .01 for  $\Delta$ CFI. This cutoff has been shown to perform well with the Mplus calculation of CFI and under different conditions of invariance and non-invariance (Chen, 2007).

## Results

### Readability Estimates

Table 2 presents the four readability estimates by US grade-level, the average across the four indices, and the reading age in years. Estimates were calculated for the introductory text, individual items, subscales and total scale. The introductory text was found to have a reading age considerably greater than 11. Similarly, items 3, 13, 16 (emotional), 4, 20 (prosocial), 10, 15 (hyperactivity), and 14 (peer problems) were calculated as having readability estimates greater than 12 years old. Of the five subscales, emotional problems and hyperactivity were calculated as having the highest reading ages ( $>12$ ). However, despite appropriate estimates for the remaining

subscales and total scale, conduct problems was the only subscale not to include any items with a reading age greater than 12 years. Items 10, 13, 15, 16 and 20 were of particular concern with reading ages greater than 15 years.

### **Item Quality Criteria**

The measure's items, floor/ceiling effects, and quality scores can be found in Table 2. While we expected varied quality, results were not favorable with 17 items (68%) shown to have poor item quality (see Table 2). Specifically, of the SDQ's 25 items, 14 (four emotional problems, four conduct problems, three hyperactivity-inattention, two prosocial, and one peer problems) clearly include more than one statement, and therefore request a response about more than one experience. The measure also has five reversed items across the conduct problems, hyperactivity, and peer problems scales. All 20 difficulties items showed substantial floor effects, ranging from 21–85%, and a further eight also had ceiling effects, ranging from 15–34%. The prosocial items showed ceiling effects, ranging from 29–69%, and one also had a floor effect at 16%.

### **Readability vs. Item Quality.**

Though our readability methodology suffers from applying formulas to short texts (Oakland & Lane, 2004), this was considered alongside item quality criteria, so that items could be evaluated more comprehensively. For instance, the item with the lowest reading age, "*I worry a lot*", also performed well in terms of item quality since it is not reversed, and consists of a single statement. Conversely, the item "*I fight a lot. I can make other people do what I want*" has a low reading age, but introduces confusion since respondents must affirm two independent behaviors. Another consideration is that the measure is often deployed in schools, as was the case for our sample (e.g., Wigelsworth et al., 2012). The item "*I am easily distracted, I find it*

*difficult to concentrate*” has the highest reading age because it contains several multiple syllable words. On one hand, young people in schools may regularly be talked to about concentration and therefore be more readily primed to recognize these words than readability formulas would suggest. However, item quality criteria confirm that this statement is unnecessarily complex, containing two statements. Readability and age-appropriateness of measures are therefore more complex than any one type of analysis might suggest.

### **Factor Structure and Measurement Invariance**

School year group was available for all but one participant, and missingness for SDQ responses ranged from .5-1.5%. Variance and Covariance coverage were high (>.97) for SDQ items suggesting that estimates were likely to be trustworthy (Muthén, Muthén, & Asparouhov, 2017). Since data were not missing completely at random,  $\chi^2(13289) = 17509.62, p < .0001$ , we explored missingness at the subscale level, using gender, age, ethnicity, self-reported wellbeing, special educational needs and free school meal eligibility as predictors. Special educational needs (OR = .25–.37) predicted less missing data for all subscales. Unclassified ethnicity predicted less missing data for all but the conduct problems subscale (OR = .01–.21). Asian ethnicity predicted less missing data for peer problems, prosocial behaviour and hyperactivity (OR = .28–.35). Higher wellbeing predicted less missing data for peer problems and prosocial behaviour (OR = .92–.93), while girls (OR = .33) and those from black ethnic backgrounds (OR = .28) were less likely to have missing data for prosocial behaviour.

Fit of all models estimated is provided in Table 3. The original correlated 5-Factor structure was found to have poor fit, as did the higher-order model. The bifactor structure of the four difficulties subscales similarly indicated a total difficulties score to be problematic, even though bifactor structures are highly parameterized with a tendency to overfit (Murray &



Johnson, 2013). As expected, given the flexibility of such models, the ESEM solution provided a much better fit to the data. Nevertheless, primary ESEM loadings were strongly related to their corresponding parameters in the CFA model. This was established via a correlation between loadings from the ESEM and CFA models ( $r = .65$ ) following the example by Marsh et al. (2013b).

The ESEM solution (see supplemental table S1) revealed nine items to cross-load with a discrepancy of  $< .30$  between the highest and second highest loadings, which is indicative of problems with the item (Matsunaga, 2010). Each of the five reversed items also loaded above  $.34$  on the prosocial factor, and less strongly on their theorized difficulties factors. The prosocial factor was not correlated with the emotional problems and peer problems factors at a significant level. Similarly, the hyperactivity factor was not significantly associated with the peer problems factor. Factor correlations beyond this were in expected directions, with the largest associations seen between hyperactivity and conduct problems ( $r = .49$ ), and emotional problems and peer problems ( $r = .38$ ). Sensitivity analysis also revealed that accounting for the categorical nature of items via WLSMV had little impact on results. No changes in fit or loadings were seen in terms of recommended cutoffs, supporting confidence in the main results reported based on MLR.

Acceptable model fit was found for the two age groups separately. Consistent with findings for the parent version with middle and older adolescents (He et al., 2013), but counter to our hypothesis based on previous readability evidence, approximate age measurement invariance was supported, as the  $\Delta CFI$  was found to be below  $.01$  in all comparisons (see Table 4).

### **Discussion**

Though the self-report SDQ is widely used, including to study age differences (Deighton et al., 2019; Hafekost et al., 2016; Johnston & Gowers, 2005), evidence of its age

appropriateness has been limited. Building on existing evidence (He et al., 2013; Patalay et al., 2018) we addressed this gap by considering the measure's item-level readability, item quality, factor structure, and age measurement invariance. Items showed a wide range of reading ages, which was more varied than previous subscale-level analysis had indicated (Patalay et al., 2018). Many items also appeared to be too difficult for the intended age group. Beyond this, a substantial proportion of the measure was found to be problematic in terms of item quality, and the proposed factor structure was a poor fit to the data. ESEM allowed approximate measurement invariance to be tested between students in year seven versus year nine, which suggested that this flexible structure was invariant across these groups.

While Patalay and colleagues (2018) had already demonstrated the measure may not be suitable for adolescents under 13, their analysis was unable to clarify which items might be problematic. In fact, our results suggest scale and subscale-level reading scores could be misleading since they suggested levels around age 11. Counter to our first hypothesis, item-level readability was much more varied than that found previously at the subscale level. We found some items to be much more difficult and others much easier. For instance, while the emotional problems subscale had an average reading age of 12.68, the item "*I worry a lot*" performed much better with an average reading age of 5.41. This item is therefore an example of optimal simplicity.

Beyond the item-level analysis, the instructions did not meet recommendations published elsewhere that even adult scales should have reading ages of no more than 12 (Terwee et al., 2007). This suggests there may have been problems even for higher quality items. In fact, special attention to instructions has been recommended for surveys with young people since clearer and more detailed instructions can be associated with greater reliability (Omran,

Wakefield-Scurr, Smith, & Brown, 2018). Similarly, Though the stated reference period in the SDQ instructions is clear, i.e. not subjective such as “often”, but finite, “over the last six months”, this may not be appropriate to the assessment of symptoms in adolescents. Younger adolescents, in particular, tend to find long reference periods challenging, and guidelines suggest very recent or current reference periods may lead to more valid responses in this age group (Bell, 2007; de Leeuw, 2011).

As well as clarifying readability analysis, consideration of item quality criteria revealed the measure to have certain other problems. Alongside the fact that over half of items contain multiple statements, the SDQ also contains five reversed items. While such items are common in scale development, it is generally advised that these be avoided since they tend not to factor well with other constructs or be opposite indicators as developers intend them to be (Ebesutani et al., 2012; Suárez-Alvarez et al., 2018; van Sonderen et al., 2013). In the current study it was clear the reversed items were not measuring the subscale constructs cleanly, as ESEM results revealed all these items to have substantial cross-loadings. This is also consistent with findings in other language versions of the SDQ (Garrido et al., 2018; van de Looij-Jansen, Goedhart, de Wilde, & Treffers, 2011). Specifically, we found each of the reversed items loaded more strongly on the prosocial factor than on their respective theorized factors. Some shared variance could reasonably be anticipated. However, the magnitude of these cross loadings (particularly on the prosocial factor), suggests that beyond age-appropriateness, these items may also face wider validity problems. Reversed items can affect instrument structure through misresponse since their content may not be perceived as opposite to positively worded statements (Weijters & Baumgartner, 2012). Though we did not explicitly examine common method effects, our ESEM

results suggest reversed items could have introduced noise into the structure through similarity to prosocial items, as they all relate to positive behaviors.

Item quality criteria also provided insight into the measure's applicability across the range of symptoms. In our community sample, which showed above average levels of mental health difficulties, high levels of floor or ceiling effects were seen for every item. While this is a common feature of clinical measures used in samples with predominantly healthy individuals, the measure's use may be somewhat limited, particularly if recommended dimensional approaches to understanding symptoms are adopted (Krueger et al., 2018). This is because measures with high floor and ceiling effects tend to have less discriminatory ability and responsiveness; in other words they may be less able to detect change and discriminate between individuals with different levels of problems (e.g., high versus borderline; de Vet, Terwee, Mokkink, & Knol, 2011). The three-point response format may contribute to the skewed nature of the data since having more categories can be associated with higher reliability and validity (Lozano, García-Cueto, & Muñiz, 2008). While there is relatively little research on number of response categories with young people, available evidence suggests around four options may provide a good balance in terms of memory, reading, reliability and stability (Bell, 2007; Omrani et al., 2018).

Beyond the issues already identified, further elements have also been suggested as indicators of psychometric quality. Of particular relevance to the current study, is that measures should ideally be developed in consultation with the target population (Irwing & Hughes, 2018; Terwee et al., 2007), since this allows assessment of acceptability and bias of items. It is possible that some of the psychometric problems identified in the SDQ are compounded by such issues,

as to the authors' knowledge, such consultation did not take place in the development of the SDQ.

Regarding the SDQ's structure, we found the five correlated subdomains to be a poor fit to the data, and uncovered substantial shared variance across factors in the ESEM solution. Both the higher-order internalizing/externalizing model, and the bifactor difficulties model also failed to show good fit. These results indicate that using the SDQ to calculate subdomain scores is questionable (Raykov & Marcoulides, 2011). Our ESEM results further suggest the hypothesized structure may be problematic since several items loaded onto more than one factor.

The instrument's poor fit may also be explained by satisficing theory, which is considered to be of particular relevance to adolescents (Krosnick, 1991; Omrani et al., 2018). This holds that the greater the cognitive demand on participants, the lower the reliability of their responses, as steps involved in providing appropriate responses are skipped (Krosnick, 1991; Omrani et al., 2018). The following results in this study could support such an account: 1) subscales showed mixed reliability, as measured through internal consistency; 2) the instructions had a higher reading age than the lowest limit of the intended population; 3) many items did not have appropriate reading ages, with some at very high levels; 4) the reference period of six months is often considered to be inappropriate for younger adolescents (Bell, 2007; de Leeuw, 2011); 5) several items, particularly those with reverse wording, were found to tap into more than one construct; 6) many items contained multiple statements which tend to increase cognitive load (Oakland & Lane, 2004).

Since we found the hypothesized CFA structures to be inadequate, we proceeded to invariance testing with the ESEM model, which as expected showed excellent fit. We found no evidence of differences in how 11–12-year-olds versus 13–15-year-olds responded using this

flexible model. Since we used  $\Delta CFI$  to establish approximate invariance, we interpret our findings as suggesting that any differences between groups are likely insubstantial. Though we anticipated older students might respond markedly differently, as previous research suggested the SDQ may be more appropriate to their reading ability, (Patalay et al., 2018), our results suggest that both groups responded to it with the same level of ease and/or difficulty. Still, our readability evidence suggests that items with a reading age above 14 may have been too difficult for both groups. In fact our sample had below average ability in reading which could also support the idea that approximate invariance was caused in part by high reading age items being equally difficult for both groups. Further work is needed (e.g. cognitive interviews with young people) to consolidate our findings.

Taken together, our findings indicate a large proportion of self-report SDQ items are less appropriate for use with younger populations. The current study is the first to provide a detailed item-level readability analysis, thus uncovering specific issues with the self-report SDQ. While previous evidence suggested four of the five subscales had reading ages higher than the recommended minimum age (Patalay et al., 2018), the current study indicates this may be not be the case for *all* items. Still, our findings call for caution when using the self-report SDQ with younger adolescents or populations with mental health difficulties, since this group may have below average reading ability (Jensen et al., 2006; Moilanen, Shaw, & Maxwell, 2010). It should also be noted that self-report adolescent mental health measures have generally been found to be poor in terms of psychometric quality (Bentley et al., 2019). It is therefore important that researchers and clinicians consider carefully the psychometric quality and reading age of their chosen instrument in relation to their sample (Jensen et al., 2006).

Our study brought together robust and complementary methodological approaches to comprehensively assess age-appropriateness of a widely used measure for the first time. Indeed our findings highlight the importance of conducting supplementary analysis such as readability and item quality alongside invariance testing, since these can provide additional insight. Together, assessment of item quality and readability with factor analysis suggested that the scale contains several difficult statements and psychometrically poor items with a response scale that prevents it from capturing the full spectrum of symptoms experienced in the general population (Terwee et al., 2007).

Despite these methodological strengths, a number of limitations must be acknowledged. First, though we attempted to overcome the problem of losing information about items when applying readability formulas to subscales, our item-level readability results should be interpreted carefully. These formulas were not designed for this purpose and therefore may not be as reliable as when used with longer passages (Oakland & Lane, 2004). However, we are confident that high-scoring items are likely inappropriate for younger audiences since they also showed poor item quality. It has also been suggested that assessment of readability at the item level is vital since this reflects how respondents actually perceive scale texts, particularly since individual items may be skipped or invalid responses provided when demands are too great (Calderón, Morales, Liu, & Hays, 2006). In addition, although readability results were considered alongside other well-established indicators of item quality, these were not based on a standardized measure.

We also treated items as continuous so we could employ the more robust  $\Delta$ CFI index for invariance testing, though our data were ordinal. The skewness in our data was controlled for by using MLR and sensitivity analysis using WLSMV supported these findings. Thirdly, though our

large sample size was likely an asset for assessing the generalizability of floor and ceiling effects, and the factor structure of the measure, it is not currently clear how approximate difference testing using  $\Delta CFI$  is affected by samples of the magnitude reported here. It is also possible that the explanations provided via the online portal affected measurement invariance by masking the differences in ability between the older and younger cohort. However, in any large-scale research with young people it is likely that support would be provided in some form (e.g. by a teacher or researcher). It is therefore likely very difficult to provide measurement invariance analysis across age groups without some kind of confound for ability.

Results must also be interpreted only for the ESEM model, which is less restrictive, with cross-loadings freely estimated. The theorized CFA model by Goodman et al. (1989) was not suitable for measurement invariance testing, and we therefore stress that invariance of this model could not be determined. Though lack of control over a priori structure in ESEM is therefore a limitation (Marsh, Liem, Martin, Morin, & Nagengast, 2011), five factors corresponding to the original theoretical model were extracted in order to accommodate issues such as cross-loadings without resorting to post-hoc model modification. Similarly, though the large number of parameters in ESEM is a limitation, our large sample size was likely able to handle this with a ratio of 163.7 cases per parameter. Finally, though our sample was large, it was not representative of the general population since deprivation was seen at higher levels, given the focus of the project from which data were drawn.

### **Conclusion and Future Directions**

While the self-report SDQ has been used extensively, our study suggests the measure would benefit from revisions three decades on from its original development. It is perhaps surprising that such a widely used measure suffers from issues such as those described here,



although as our findings suggest, this is possibly due to the lack of attention to robust scale development practices (e.g. omission of cognitive interviews with young people). Items should be simplified, with reversed wording and multiple statements replaced with simpler alternatives, and more straightforward language used for items with high reading ages. We also recommend that such amendments be made in consultation with young people in line with policy and psychometric best practice (Deighton et al., 2014; Irwing & Hughes, 2018; Terwee et al., 2007).

## References

- Asparouhov, T., & Muthén, B. (2018). *SRMR in Mplus*. Retrieved from <http://www.statmodel.com/download/SRMR2.pdf>
- Bell, A. (2007). Designing and testing questionnaires for children. *Journal of Research in Nursing, 12*(5), 461-469. doi:10.1177/1744987107079616
- Bentley, N., Hartley, S., & Bucci, S. (2019). Systematic Review of Self-Report Measures of General Mental Health and Wellbeing in Adolescent Mental Health. *Clinical Child Family Psychology Review*. doi:10.1007/s10567-018-00273-x
- Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): Factor Structure and Gender Equivalence in Norwegian Adolescents. *PLOS ONE, 11*(5), e0152202. doi:10.1371/journal.pone.0152202
- Brown, T. A. (2015). *Confirmatory factor analysis for applied research*: Guilford Publications.
- Calderón, J. L., Morales, L. S., Liu, H., & Hays, R. D. (2006). Variation in the readability of items within surveys. *American journal of medical quality : the official journal of the American College of Medical Quality, 21*(1), 49-56. doi:10.1177/1062860605283572
- Chall, J. S., & Dale, E. (1995). *Readability revisited : the new Dale-Chall readability formula.*: Brookline Books.
- Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. doi:10.1080/10705510701301834
- Chen, F. F., & Zhang, Z. (2018). Bifactor Models in Psychometric Test Development *The Wiley Handbook of Psychometric Testing*.

- Cheung, G. W., & Rensvold, R. B. (2002). Evaluating Goodness-of-Fit Indexes for Testing Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 233-255. doi:10.1207/S15328007SEM0902\_5
- Chiorri, C., Hall, J., Casely-Hayford, J., & Malmberg, L.-E. (2016). Evaluating Measurement Invariance Between Parents Using the Strengths and Difficulties Questionnaire (SDQ). 23(1), 63-74. doi:10.1177/1073191114568301
- Coleman, M., & Liao, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2), 283-284. doi:10.1037/h0076540
- Dale, E., & Chall, J. S. (1948). A Formula for Predicting Readability: Instructions. *Educational Research Bulletin*, 27(2), 37-54.
- de Leeuw, E. D. (2011). *Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting*. Retrieved from Finland:  
[http://www.aka.fi/globalassets/awanhat/documents/tiedostot/lapset/presentations-of-the-annual-seminar-10-12-may-2011/surveying-children-and-adolescents\\_de-leeuw.pdf](http://www.aka.fi/globalassets/awanhat/documents/tiedostot/lapset/presentations-of-the-annual-seminar-10-12-may-2011/surveying-children-and-adolescents_de-leeuw.pdf)
- de Vet, H. C., Terwee, C. B., Mokkink, L. B., & Knol, D. L. (2011). *Measurement in medicine: a practical guide*: Cambridge University Press.
- Deighton, J., Croudace, T., Fonagy, P., Brown, J., Patalay, P., & Wolpert, M. (2014). Measuring mental health and wellbeing outcomes for children and adolescents to inform practice and policy: a review of child self-report measures. *Child and Adolescent Psychiatry and Mental Health*, 8(1), 14. doi:10.1186/1753-2000-8-14
- Deighton, J., Lereya, S. T., Casey, P., Patalay, P., Humphrey, N., & Wolpert, M. (2019). Prevalence of mental health problems in schools: poverty and other risk factors among 28

000 adolescents in England. *The British Journal of Psychiatry*, 1-3.

doi:10.1192/bjp.2019.19

Department for Education. (2014). *Statistical First Release National curriculum assessments at key stage 2 in*

*England, 2014 (Revised)*. Retrieved from

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/428838/SFR50\\_2014\\_Text.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/428838/SFR50_2014_Text.pdf)

Department for Education. (2016). *National curriculum assessments at key stage 2 in England, 2016 (revised)*. Retrieved from

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/577296/SFR62\\_2016\\_text.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/577296/SFR62_2016_text.pdf)

Department for Education. (2017a). Pupil premium: allocations and conditions of grant 2016 to 2017. Retrieved from <https://www.gov.uk/government/publications/pupil-premium-conditions-of-grant-2016-to-2017>

Department for Education. (2017b). Schools, pupils and their characteristics: January 2017.

Retrieved from

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/650547/SFR28\\_2017\\_Main\\_Text.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/650547/SFR28_2017_Main_Text.pdf)

Department for Education. (2017c). *Special educational needs in England: January 2017*.

Retrieved from

[https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment\\_data/file/633031/SFR37\\_2017\\_Main\\_Text.pdf](https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/633031/SFR37_2017_Main_Text.pdf)

- Dray, J., Bowman, J., Freund, M., Campbell, E., Hodder, R. K., Lecathelinais, C., & Wiggers, J. (2016). Mental health problems in a regional population of Australian adolescents: association with socio-demographic characteristics. *Child and Adolescent Psychiatry and Mental Health, 10*(1), 32. doi:10.1186/s13034-016-0120-9
- Ebesutani, C., Drescher, C. F., Reise, S. P., Heiden, L., Hight, T. L., Damon, J. D., & Young, J. (2012). The Loneliness Questionnaire–Short Version: An Evaluation of Reverse-Worded and Non-Reverse-Worded Items Via Item Response Theory. *Journal of Personality Assessment, 94*(4), 427-437. doi:10.1080/00223891.2012.662188
- Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science, 8*(4), 370-378. doi:10.1177/1948550617693063
- Gadermann, A. M., Guhn, M., & Zumbo, B. D. (2012). Estimating ordinal reliability for Likert-type and ordinal item response data: a conceptual, empirical, and practical guide. *Practical Assessment, Research & Evaluation, 17*(3).
- Garrido, L. E., Barrada, J. R., Aguasvivas, J. A., Martínez-Molina, A., Arias, V. B., Golino, H. F., . . . Rojo-Moreno, L. (2018). Is Small Still Beautiful for the Strengths and Difficulties Questionnaire? Novel Findings Using Exploratory Structural Equation Modeling. *Assessment, 1073191118780461*. doi:10.1177/1073191118780461
- Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). 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, 38*(8), 1179-1191. doi:10.1007/s10802-010-9434-x

Goodman, R. (2001). Psychometric Properties of the Strengths and Difficulties Questionnaire.

*Journal of the American Academy of Child & Adolescent Psychiatry*, 40(11), 1337-1345.

doi:<https://doi.org/10.1097/00004583-200111000-00015>

Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A

pilot study on the validity of the self-report version. *European Child & Adolescent*

*Psychiatry*, 7(3), 125-130. doi:10.1007/s007870050057

Gunning, F. (1952). *The technique of clear writing*: McGraw-Hill.

Hafekost, J., Lawrence, D., Boterhoven de Haan, K., Johnson, S. E., Saw, S., Buckingham, W.

J., . . . Zubrick, S. R. (2016). Methodology of Young Minds Matter: The second

Australian Child and Adolescent Survey of Mental Health and Wellbeing. *Australian &*

*New Zealand Journal of Psychiatry*, 50(9), 866-875. doi:10.1177/0004867415622270

He, J.-P., Burstein, M., Schmitz, A., & Merikangas, K. R. J. J. o. A. C. P. (2013). The Strengths

and Difficulties Questionnaire (SDQ): the Factor Structure and Scale Validation in U.S.

Adolescents. *41*(4), 583-595. doi:10.1007/s10802-012-9696-6

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis:

Conventional criteria versus new alternatives. *Structural Equation Modeling: A*

*Multidisciplinary Journal*, 6(1), 1-55. doi:10.1080/10705519909540118

Humphrey, N., & Wigelsworth, M. (2016). Making the case for universal school-based mental

health screening. *Emotional and Behavioural Difficulties*, 21(1), 22-42.

doi:10.1080/13632752.2015.1120051

Irwing, P., & Hughes, D. J. (2018). Test Development. In P. Irwing, T. Booth, & D. J. Hughes

(Eds.), *The Wiley Handbook of Psychometric Testing* (pp. 3-47).

- Jackson, R. H., Davis, T. C., Bairnsfather, L. E., George, R. B., Crouch, M. A., & Gault, H. (1991). Patient reading ability: an overlooked problem in health care. *Southern medical journal*, *84*(10), 1172-1175. doi:10.1097/00007611-199110000-00004
- Janan, D., & Wray, D. (2012). *Readability: the limitations of an approach through formulae*. Paper presented at the British Educational Research Association Annual Conference, University of Manchester. <http://www.leeds.ac.uk/educol/documents/213296.pdf>
- Jensen, S. A., Fabiano, G. A., Lopez-Williams, A., & Chacko, A. (2006). The reading grade level of common measures in child and adolescent clinical psychology. *Psychological Assessment*, *18*(3), 346-352. doi:10.1037/1040-3590.18.3.346
- Johnston, C., & Gowers, S. (2005). Routine Outcome Measurement: A Survey of UK Child and Adolescent Mental Health Services. *Child and Adolescent Mental Health*, *10*(3), 133-139. doi:doi:10.1111/j.1475-3588.2005.00357.x
- Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Retrieved from <https://stars.library.ucf.edu/cgi/viewcontent.cgi?article=1055&context=istlibrary>
- Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, *5*(3), 213-236. doi:10.1002/acp.2350050305
- Krueger, R. F., Kotov, R., Watson, D., Forbes, M. K., Eaton, N. R., Ruggero, C. J., . . . Zimmermann, J. (2018). Progress in achieving quantitative classification of psychopathology. *World Psychiatry*, *17*(3), 282-293. doi:10.1002/wps.20566

- Lai, K., & Green, S. B. (2016). The Problem with Having Two Watches: Assessment of Fit When RMSEA and CFI Disagree. *Multivariate Behavioral Research, 51*(2-3), 220-239. doi:10.1080/00273171.2015.1134306
- Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936-949. doi:10.3758/s13428-015-0619-7
- Lozano, L. M., García-Cueto, E., & Muñiz, J. (2008). Effect of the Number of Response Categories on the Reliability and Validity of Rating Scales. *Methodology, 4*(2), 73-79. doi:10.1027/1614-2241.4.2.73
- Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological Measurement Fruitfulness of Exploratory Structural Equation Modeling (ESEM): New Approaches to Key Substantive Issues in Motivation and Engagement. *Journal of Psychoeducational Assessment, 29*(4), 322-346. doi:10.1177/0734282911406657
- Marsh, H. W., Nagengast, B., & Morin, A. J. S. (2013a). Measurement invariance of big-five factors over the life span: ESEM tests of gender, age, plasticity, maturity, and la dolce vita effects. *Developmental Psychology, 49*(6), 1194-1218. doi:10.1037/a0026913
- Marsh, H. W., Vallerand, R. J., Lafrenière, M.-A. K., Parker, P., Morin, A. J. S., Carbonneau, N., . . . Paquet, Y. (2013b). Passion: Does one scale fit all? Construct validity of two-factor passion scale and psychometric invariance over different activities and languages. *Psychological assessment, 25*(3), 796-809. doi:10.1037/a0032573



Matsunaga, M. (2010). How to factor-analyze your data right: do's, don'ts, and how-to's.

*International journal of psychological research*, 3(1), 97-110.

doi:<https://doi.org/10.21500/20112084.854>

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, 93(3), 568-592. doi:10.1037/0021-9010.93.3.568

Moilanen, K. L., Shaw, D. S., & Maxwell, K. L. (2010). Developmental cascades: Externalizing, internalizing, and academic competence from middle childhood to early adolescence.

*Development and Psychopathology*, 22(3), 635-653. doi:10.1017/S0954579410000337

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher-order models of human cognitive ability structure. *Intelligence*, 41(5), 407-422. doi:<https://doi.org/10.1016/j.intell.2013.06.004>

Muthén, B., & Muthén, L. (2016). warning about parameters and clusters. Retrieved from

<http://www.statmodel.com/discussion/messages/12/20967.html?1463144022>

Muthén, B. O., Muthén, L. K., & Asparouhov, T. (2017). *Regression and mediation analysis using Mplus*: Muthén & Muthén Los Angeles, CA.

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide*. Eighth Edition. Los Angeles, CA: Muthén & Muthén.

Ng Fat, L., Scholes, S., Boniface, S., Mindell, J., & Stewart-Brown, S. (2017). Evaluating and establishing national norms for mental wellbeing using the short Warwick–Edinburgh Mental Well-being Scale (SWEMWBS): findings from the Health Survey for England. *Quality of Life Research*, 26(5), 1129-1144. doi:10.1007/s11136-016-1454-8

- NHS Digital. (2018). *Mental Health of Children and Young People in England, 2017 Summary of key findings*. Retrieved from <https://files.digital.nhs.uk/F6/A5706C/MHCYP%202017%20Summary.pdf>.
- Oakland, T., & Lane, H. B. (2004). Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests. *International Journal of Testing*, 4(3), 239-252. doi:10.1207/s15327574ijt0403\_3
- Omrani, A., Wakefield-Scurr, J., Smith, J., & Brown, N. (2018). Survey Development for Adolescents Aged 11–16 Years: A Developmental Science Based Guide. *Adolescent Research Review*. doi:10.1007/s40894-018-0089-0
- Ortuño-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., Riba, S. S. i., & Muñiz, J. (2015a). The assessment of emotional and Behavioural problems: Internal structure of The Strengths and Difficulties Questionnaire. *International Journal of Clinical and Health Psychology*, 15(3), 265-273. doi:<https://doi.org/10.1016/j.ijchp.2015.05.005>
- Ortuño-Sierra, J., Fonseca-Pedrero, E., Aritio-Solana, R., Velasco, A. M., de Luis, E. C., Schumann, G., . . . consortium, I. (2015b). New evidence of factor structure and measurement invariance of the SDQ across five European nations. *European Child & Adolescent Psychiatry*, 24(12), 1523-1534. doi:10.1007/s00787-015-0729-x
- Panayiotou, M., Humphrey, N., & Wigelsworth, M. (2019). An empirical basis for linking social and emotional learning to academic performance. *Contemporary Educational Psychology*, 56, 193-204. doi:<https://doi.org/10.1016/j.cedpsych.2019.01.009>
- Patalay, P., Hayes, D., & Wolpert, M. (2018). Assessing the readability of the self-reported Strengths and Difficulties Questionnaire. *BJPsych Open*, 4(2), 55-57. doi:10.1192/bjo.2017.13

- Percy, A., McCrystal, P., & Higgins, K. (2008). Confirmatory Factor Analysis of the Adolescent Self-Report Strengths and Difficulties Questionnaire. *24*(1), 43-48. doi:10.1027/1015-5759.24.1.43
- Polanczyk, G. V., Salum, G. A., Sugaya, L. S., Caye, A., & Rohde, L. A. (2015). Annual Research Review: A meta-analysis of the worldwide prevalence of mental disorders in children and adolescents. *Journal of Child Psychology and Psychiatry*, *56*(3), 345-365. doi:10.1111/jcpp.12381
- Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*: Routledge.
- Raykov, T., & Marcoulides, G. A. (2016). Scale Reliability Evaluation Under Multiple Assumption Violations. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*(2), 302-313. doi:10.1080/10705511.2014.938597
- Saris, W. E. (2014). *Design, evaluation, and analysis of questionnaires for survey research* (Second edition. ed.). Hoboken: Wiley.
- Sass, D. A., Schmitt, T. A., & Marsh, H. W. (2014). Evaluating Model Fit With Ordered Categorical Data Within a Measurement Invariance Framework: A Comparison of Estimators. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(2), 167-180. doi:10.1080/10705511.2014.882658
- Suárez-Alvarez, J., Pedrosa, I., Lozano Fernández, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema*, *30*(2), 149-158. doi:10.7334/psicothema2018.33
- Terwee, C. B., Bot, S. D. M., de Boer, M. R., van der Windt, D. A. W. M., Knol, D. L., Dekker, J., . . . de Vet, H. C. W. (2007). Quality criteria were proposed for measurement

- properties of health status questionnaires. *Journal of Clinical Epidemiology*, 60(1), 34-42.  
doi:<https://doi.org/10.1016/j.jclinepi.2006.03.012>
- Tóth-Király, I., Bőthe, B., Rigó, A., & Orosz, G. (2017). An Illustration of the Exploratory Structural Equation Modeling (ESEM) Framework on the Passion Scale. 8(1968).  
doi:10.3389/fpsyg.2017.01968
- van de Looij-Jansen, P. M., Goedhart, A. W., de Wilde, E. J., & Treffers, P. D. A. (2011). Confirmatory factor analysis and factorial invariance analysis of the adolescent self-report Strengths and Difficulties Questionnaire: How important are method effects and minor factors? *British Journal of Clinical Psychology*, 50(2), 127-144.  
doi:10.1348/014466510x498174
- van Sonderen, E., Sanderman, R., & Coyne, J. C. (2013). Ineffectiveness of Reverse Wording of Questionnaire Items: Let's Learn from Cows in the Rain. *PLOS ONE*, 8(7), e68967.  
doi:10.1371/journal.pone.0068967
- Vostanis, P. (2006). Strengths and Difficulties Questionnaire: Research and clinical applications. *Current Opinion in Psychiatry*, 19(4), 367-372.  
doi:10.1097/01.yco.0000228755.72366.05
- Weijters, B., & Baumgartner, H. (2012). Misresponse to Reversed and Negated Items in Surveys: A Review. *Journal of Marketing Research*, 49(5), 737-747.  
doi:10.1509/jmr.11.0368
- Wigelsworth, M., Humphrey, N., & Lendrum, A. (2012). A national evaluation of the impact of the secondary social and emotional aspects of learning (SEAL) programme. *Educational Psychology*, 32(2), 213-238. doi:10.1080/01443410.2011.640308

Wolpert, M., Cheng, H., & Deighton, J. (2015). Measurement Issues: Review of four patient reported outcome measures: SDQ, RCADS, C/ORS and GBO – their strengths and limitations for clinical use and service evaluation. *20*(1), 63-70.

doi:doi:10.1111/camh.12065

Table 1.

*SDQ subscale reliability coefficients*

	Cronbach's $\alpha$	Ordinal $\alpha$	McDonald's $\omega$ [95% CI]
Emotional problems	.74	.81	.74 [.73, .74]
Conduct problems	.64	.76	.66 [.65, .66]
Hyperactivity	.74	.80	.75 [.74, .75]
Peer problems	.59	.72	.60 [.59, .61]
Prosocial	.69	.78	.69 [.68, .69]
Total difficulties	.81	.86	.87 <sup>h</sup> [.87, .87]

*Note.* <sup>h</sup> Hierarchical omega coefficient

Table 2.

*SDQ Items Floor/Ceiling Effects, Readability Estimates by US Grade-level, Average Estimate Across Indices and Reading Age*

Floor– Ceiling %	Score	Instructions and Items	US-Grade Level				Average US Grade Level	Age
			DC	FK	GFI	CLI		
		For each item, please mark the box for Not True, Somewhat True or Certainly True. It would help us if you answered all items as best you can even if you are not absolutely certain or the item seems daft! Please give your answers on the basis of how things have been for you over the last six months.	6.49	6.8	9.22	6.21	7.20	13.20†
		<b>Emotional problems</b>	<b>6.86</b>	<b>5.5</b>	<b>9.90</b>	<b>4.43</b>	<b>6.68</b>	<b>12.68†</b>
43.7–18.4	-	3) I get a lot of headaches, stomach-aches or sickness	5.84	4.9	8.04	7.69	6.62	12.62†
				1				
31.1–29.6	+	8) I worry a lot	3.83	0.7	1.60	-8.51	-0.59	5.41
				2				
55.6–11.7	-	13) I am often unhappy, down-hearted or tearful	6.24	9.0	14.23	9.36	9.73	15.73†
				9				

SDQ AGE APPROPRIATENESS

27.3–30.7	-	16) I am <u>nervous</u> in new situations. I easily lose confidence	10.20	7.6	14.00	5.32	9.28	15.28†
				0				
50.9–14.6	-	24) I have many fears, I am easily scared	6.01	3.8	8.20	1.81	4.96	10.96
				1				
		<b>Conduct problems</b>	<b>4.69</b>	<b>2.8</b>	<b>4.08</b>	<b>1.45</b>	<b>3.26</b>	<b>9.26</b>
				<b>2</b>				
39.4–23	-	5) I get very angry and often lose my temper	4.08	4.9	3.60	2.47	3.77	9.77
				1				
39.5–7.4	-	7) I usually do as I am told (R)	6.24	4.0	8.51	-4.08	3.67	9.67
				1				
74.7–5	-	12) I fight a lot. I can make other people do what I want	3.96	0.5	2.60	-2.27	1.20	7.20
				2				
56.6–14	-	18) I am often <u>accused of</u> lying or cheating	6.01	5.2	3.20	4.01	4.61	10.61
				3				
82.7–4	+	22) I take things that are not mine from home, school or elsewhere	4.23	1.8	4.80	6.23	4.27	10.27
				3				
		<b>Hyperactivity</b>	<b>7.06</b>	<b>4.6</b>	<b>9.56</b>	<b>4.96</b>	<b>6.56</b>	<b>12.56†</b>
				<b>6</b>				
21.1–34.1	-	2) I am <u>restless</u> , I cannot stay still for long	5.84	2.3	3.60	3.12	3.72	9.72
				2				



SDQ AGE APPROPRIATENESS

35.7–27	+	10) I am constantly <u>fidgiting</u> or squirming	11.83	8.3	15.73	11.60	11.88	17.88†
				4				
26.3–30.6	-	15) I am easily distracted, I find it difficult to <u>concentrate</u>	10.45	9.5	20.00	9.46	12.36	18.36†
				5				
28.7–15.1	-	21) I think before I do things (R)	3.93	0.5	2.40	-0.16	1.68	7.68
				6				
31.4–13	-	25) I finish the work I'm doing. My <u>attention</u> is good (R)	3.88	2.8	6.00	1.20	3.49	9.49
				8				
		<b>Peer problems</b>	<b>4.99</b>	<b>4.1</b>	<b>5.75</b>	<b>2.85</b>	<b>4.43</b>	<b>10.43</b>
				<b>3</b>				
63.8–10	-	6) I am usually on my own. I generally play alone or keep to myself	5.11	5.6	8.51	0.96	5.06	11.06
				7				
85.1–2.8	-	11) I have one good friend or more (R)	3.98	1.0	2.80	0.12	1.46	7.46
				6				
40–9	-	14) Other people my age generally like me (R)	3.98	7.3	8.51	6.00	6.46	12.46†
				2				
70.3–8.2	+	19) Other children or young people pick on me or bully me	5.62	4.7	4.40	4.49	4.81	10.81
				5				
44.5–14	+	23) I get on better with adults than with people my own age	5.55	3.8	4.80	3.29	4.37	10.37
				4				

		<b>Prosocial</b>	<b>5.35</b>	<b>5.2</b>	<b>4.87</b>	<b>5.36</b>	<b>5.20</b>	<b>11.20</b>
				<b>3</b>				
2.9–59.8	-	1) I try to be nice to other people. I care about their feelings	5.17	2.4	2.60	1.35	2.88	8.88
				<b>0</b>				
8.6–47.6	+	4) I usually share with others (food, games, pens etc.)	7.59	8.9	8.04	6.39	7.74	13.74†
				<b>2</b>				
5.6–51	-	9) I am helpful if someone is hurt, upset or feeling ill	4.18	4.7	4.40	3.96	4.32	10.32
				<b>5</b>				
4.9–69	+	17) I am kind to younger children	3.93	2.4	2.40	2.78	2.89	8.89
				<b>4</b>				
16.4–28.7	+	20) I often <u>volunteer</u> to help others (parents, teachers, children)	5.84	8.9	8.04	13.57	9.09	15.09†
				<b>2</b>				
		<b>Total Scale (without instructions)</b>	<b>5.68</b>	<b>4.3</b>	<b>6.55</b>	<b>3.72</b>	<b>5.08</b>	<b>11.08</b>
				<b>5</b>				
		<b>Total Scale (with instructions)</b>	<b>5.78</b>	<b>4.3</b>	<b>6.54</b>	<b>4.24</b>	<b>5.22</b>	<b>11.22</b>
				<b>1</b>				

*Note.* In bold are the estimates for the subscales and total scale. Underlined words are those for which additional explanations were provided when the mouse was hovered over them in the online administration. + = high quality; - = low quality. (R) = Reversed items. † = items and scales with readability age above 11. DC = Dale-Chall Readability Formula; FK = Flesch-Kincaid Reading Grade; GFI = Gunning Fog Index; CLI = Coleman Liau Index.

Table 3.

*Model Fit for Main and Sensitivity Analysis Models*

Model	Estimator	$\chi^2$ (df)	AIC	BIC	RMSEA [90% CI]	CFI	TLI	SRMR	$\lambda$	$h^2$
5-Factor correlated	MLR	27966.58 (265)**	1357768.55	1358475.63	.059 [.058, .059]	.80 7	.78 1	.063	.351-.716	.123-.513
	WLSMV	44612.71 (265)**	-	-	.074 [.074, .075]	.83 2	.81 0	.081	.480-.843	.221-.710
4-Factor higher- order	MLR	19841.26 (165)**	1111239.31	1111779.98	.063 [.062, .064]	.82 2	.79 5	.057	.323-.967	.104-.488
	WLSMV	28588.03 (165)**	-	-	.075 [.075, .076]	.86 7	.84 7	.071	.461-.983	.213-.726
4-Factor bifactor	MLR	19087.85 (150)**	1109992.68	1110658.12	.065 [.064, .065]	.82 9	.78 3	.069	-.086-.707	.146-.635
	WLSMV	36973.36 (150)**	-	-	.090 [.089, .091]	.82 8	.78 2	.080	.171-.884	.206-.666
ESEM	MLR	5791.31 (185)**	1333225.12	1334597.69	.032 [.031, .032]	.96 1	.93 7	.016	-	.189-.558

WLSMV	6514.99 (185)**	-	-	.034 [.033,	.97	.96	.016	-	.270-.694
				.034]	6	1			

---

*Note.* ESEM = exploratory structural equation modeling; MLR = robust maximum likelihood; WLSMV = weighted least square mean and variance adjusted;

AIC = Akaike information criterion; BIC = Bayesian information criterion; RMSEA = root mean square error of approximation; CFI = comparative fit index;

TLI = Tucker-Lewis Index; SRMR = standardized root mean squared residual;  $\lambda$  = factor loadings;  $h^2$  = item communalities.

\*\* $p < .01$ .

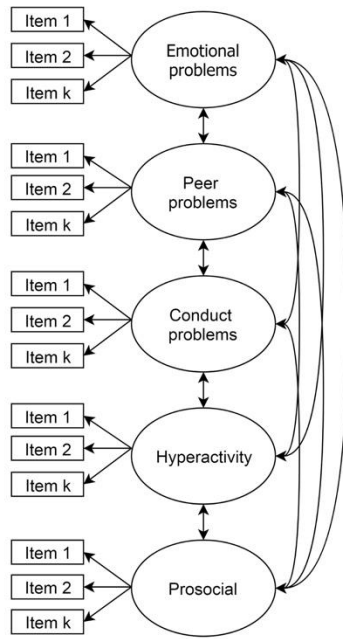
Table 4.

*ESEM Age Measurement Invariance Findings*

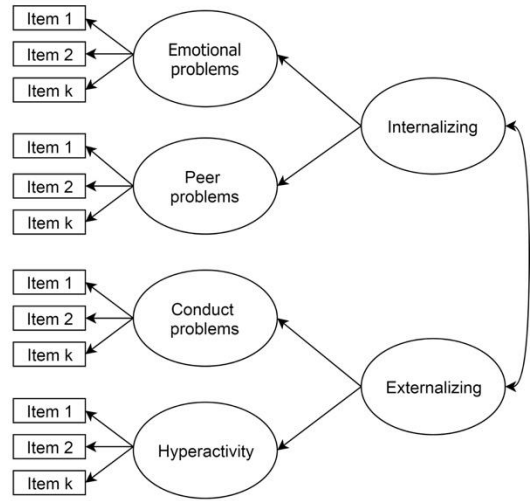
Model	$\chi^2$ (df)	AIC	BIC	RMSEA [90% CI]	CFI	TLI	SRMR	$\Delta\chi^2$ (df)	$\Delta$ CFI	$h^2$
Y7 Baseline	2779.44 (185)**	676262.55	677525.25	.030 [.029, .031]	.964	.941	.016			.194– .572
Y9 Baseline	3191.19 (185)**	653776.06	655029.61	.033 [.032, .034]	.957	.931	.018			.186– .557
Configural	5967.77 (370)**	1330038.61	1332783.73	.032 [.031, .032]	.961	.936	.017			
vs. Metric	6280.37 (470)**	1330320.63	1332233.90	.029 [.028, .029]	.959	.948	.020	409.27 (100) **	.002	
vs. Scalar	6920.98 (490)**	1330917.70	1332664.59	.029 [.029, .030]	.955	.945	.021	729.63 (20) **	.004	

*Note.* Robust maximum likelihood was used. ESEM = exploratory structural equation modeling; AIC = Akaike information criterion; BIC = Bayesian information criterion; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker-Lewis Index; SRMR = standardized root mean squared residual;  $\Delta\chi^2$  = chi-square difference test;  $\Delta$ CFI = CFI difference;  $h^2$  = item communalities.

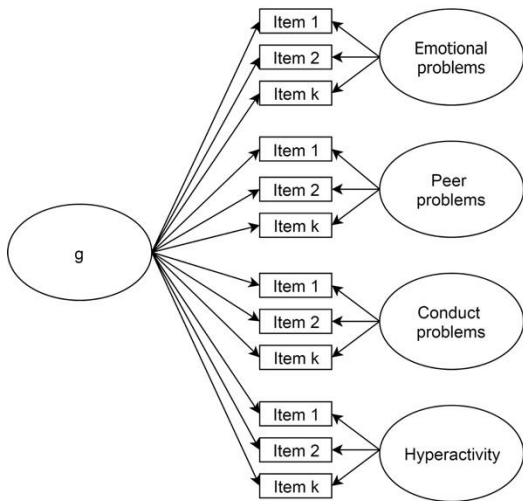
\*\* $p < .01$ .



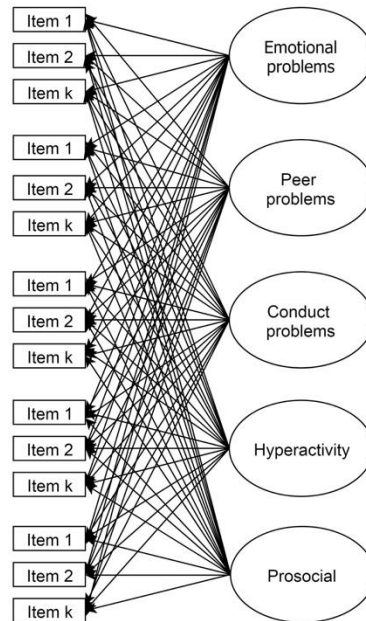
Model 1: 5-Factor correlated



Model 2: 4-Factor higher order



Model 3: 4-Factor bifactor



Model 4: 5-Factor ESEM

Figure 1. Models tested