**Running head**: Longitudinal Invariance of the SDQ

**Tracking Emotional and Behavioural Changes in Childhood: Does the Strength and Difficulties Questionnaire (SDQ) Measure the Same Constructs across Time?**

Edward M. Sosu1, Peter Schmidt2

1School of Education, University of Strathclyde, UK

2Justus Liebig University, Germany

1Correspondence should be addressed to Edward M. Sosu, School of Education, University of Strathclyde, G4 0LT, Glasgow, UK (e-mail: edward.sosu@strath.ac.uk).

Acknowledgement

The authors are grateful for the British Academy Skills acquisition award SQ120023. We also thank all the families who participated in the Growing Up in Scotland survey.

The second author`s work was also supported by Alexander von Humboldt Polish Honorary Research Fellowship.

**Abstract**

Goodman’s (1997) Strength and Difficulties Questionnaire (SDQ) is widely used to measure emotional and behavioural difficulties in childhood and adolescence. In the present study, we examined whether the SDQ measures the same construct across time, when used for longitudinal research. A nationally representative sample of parents (*N*=3375) provided data on their children at ages 4, 5, and 6. Using confirmatory factor analysis (CFA) for ordinal data, two competing models (3 versus 5-factor models) were tested to establish equivalence across time. Results showed that the 5-factor model had a superior fit to the data compared to the alternative 3-factor model which only achieved an adequate fit at a configural level. Strong longitudinal factorial invariance was established for the 5-factor parent version of the SDQ. Our findings support the use of the SDQ in longitudinal studies, and provide the important psychometric information required for basing educational, clinical and policy decisions on outcomes of the SDQ.

**Introduction**

Behavioural patterns in childhood are important predictors of future outcomes in adolescence and adulthood. For instance, significant associations have been found between behavioural difficulties in childhood and transition to school (Normandeau & Guay, 1998), academic underachievement (Masten et al., 2005), criminality, unemployment, and psychosomatic disorders (Fergusson, Horwood, & Ridder, 2005). There are also significant consequences in the form of direct aggression, stress, conflict, and high dependency for parents and professionals who interact with children identified as having behavioural problems (Hastings & Bham, 2003; Mejia & Hoglund, 2016; Neece, Green, & Baker, 2012). The associated financial costs to society are equally staggering. Estimates in the United Kingdom suggest that the state spends about ten times more on a young adult who had conduct problems as a child compared to a problem free peer (Scott, Knapp, Henderson, & Maughan, 2001). As a result, researchers, policy makers and practitioners have become interested in identifying and tracking trajectories of behavioural difficulties so that early intervention and prevention mechanisms can be deployed (Moffitt, 1993; Scottish Government, 2008; Wilson et al., 2013).

Although a wide range of scales exist, R. Goodman’s (1997) Strength and Difficulties Questionnaire (SDQ) is one of the most frequently used behavioural rating scale in educational, research and clinical settings for assessing emotional and behavioural patterns in childhood. This 25-item questionnaire measures psychological adjustment in 3-16 year olds in five sub-domains. Of these, four subscales measure difficulties (conduct problems, hyperactivity, emotional problems, and peer relational problems), while the fifth measures strengths (prosocial behaviour). There are three versions of the scale (a parent, teacher, and self-report) and these have been translated into about 79 languages (Kersten et al., 2016). Studies and reviews into the psychometric properties of the SDQ suggests that it has good structural, concurrent, discriminant, convergent, and predictive validity (e.g., Chiorri, Hall, Casely-Hayford, & Malmberg, 2016; Croft, Stride, Maughan, & Rowe, 2015; Kersten et al., 2016; Stone, Otten, Engels, Vermulst, & Janssen, 2010). While support for the 5-factor model has been established, an alternative 3-factor model which consists of an externalising (hyperactivity and conduct problems), internalising (emotional problems and peer problems) and prosocial scale has also been proposed for use due to significant conceptual overlap between subscales (Dickey & Blumberg, 2004; A. Goodman, Lamping, & Ploubidis, 2010). Other researchers have suggested eliminating ‘problematic items’ and using a version with reduced number of items (Chauvin & Leonova, 2015).

Overall, the SDQ has an advantage of brevity with completion time of about 10 minutes (McCrory & Layte, 2012). It has therefore become a key measure in several longitudinal studies funded by national governments (e.g., Bradshaw, Lewis, & Hughes, 2014; Stone et al., 2010). Crucially, the proportion of children identified as having behavioural difficulties based on outcomes of the SDQ form critical benchmarks in countries such as Scotland and the Netherlands for assessing changes in child welfare, and by extension the effectiveness of national policies (Black & Martin, 2015; Bradshaw et al., 2014; Stone et al., 2015). In the United Kingdom, schools and social workers are required to assess pupils with the SDQ and decisions about funding support and referral to mental health services are based on outcomes of SDQ scores (e.g., Department for Education, 2012; 2015). Additionally, ratings on the SDQ and changes over time are used to inform and evaluate clinical and educational interventions (Reynolds, MacKay, & Kearney, 2009; Van Sonsbeek et al., 2014). In other words, important educational, clinical, and policy decisions tend to be at least partly based on outcomes from the SDQ.

Behavioural rating instruments such as the SDQ are however subjective (Borsboom, 2005). This is crucial in longitudinal research where several factors such as age, experience, context and personal choices can change and affect how individuals respond to items (Little, 2013). Thus, the way respondents understand constructs measured by the SDQ can change over time and longitudinal studies may not accurately reflect ‘true change’ in contrast to “change due to measurement” (Oort, 2005; Widaman, Ferrer, & Conger, 2010). This phenomenon, referred to as ‘response shift’, can occur due to changes in respondents’ internal standards, their values and priorities, or their conceptualisation of a target construct across time (Oort, 2005; Schwartz & Sprangers 1999).

Schwartz and Sprangers (1999) identified three different ways in which one’s self-evaluation of a target construct can change over time. The first, *scale recalibration*, refers to changes in a respondent’s internal standards of measurement and can occur when participants revise response scale values across time. For example, a score of 2 (Certainly True) on item 5 on the SDQ (child often has temper tantrums) measured at age six may reflect a different level of tantrums than measures obtained at age four. Secondly, *reprioritization* which refers to a change in a respondent’s values can occur when participants attach a different level of importance to an item in the context of the total scale. For example, child being obedient (item 7) may have had more importance when compared to tantrums (item 5) to a parental respondent at age 4, but this prioritization may have changed over time. Finally, *reconceptualization* which refers to redefinition of a target construct can occur when there is a change in the meaning that respondents attach to items. For instance, a parent’s understanding of ‘child being restless’ (item 2) may be different when their child was four years old, compared to when they are six years of age due to the experience of being a parent over time.

Thus, for studies using the SDQ to meaningfully evaluate change in behaviour, it is important that observed changes are not due to any form of response shift on the instrument (Oort, 2005; Schwartz & Sprangers 1999; Widaman et al., 2010). These assumptions of changes in meanings of items or constructs are evaluated in this paper within a framework of invariance testing. A scale is longitudinally invariant if participants retain the same meaning of indicators and constructs across time (Little, 2013; Millsap, 2011). More specifically, invariance testing within a longitudinal framework examines the relations between constructs and the items used to measure them, and whether the relations are stable across occasions (Widaman et al., 2010). Judgement can be made about whether an instrument has strong, weak or no factorial invariance across time depending on the extent to which certain conditions are met. Assuming the structure of the model is the same across time, weak factorial invariance is achieved when all common factor loadings are the same across time, and strong invariance is reached when factor loadings, intercepts or thresholds are the same across time (Little, 2013; Meredith, 1993; Widaman et al., 2010). According to Oort (2005), the parameters tested within longitudinal measurement invariance models provide a framework for formerly evaluating different types of response shift (Table 1). Specifically, changes in the value of factor loadings of an observed item over time suggests that the item has become more or less important to the measurement of the latent construct of interest, an indication of reprioritisation. Differences in item intercepts or thresholds over time suggest that participants have changed their interpretation of response scale values, a sign that recalibration has occurred. Finally, different salient factor loadings over time or common factor loadings that change from zero to non-zero would indicate a change in the meaning attached to specific items, hence reconceptualization. Assessing whether the SDQ is invariant over time is therefore an imperative before making recommendations based on change scores derived from the instrument.

--------------------------------------------

INSERT TABLE 1 HERE

---------------------------------------------

**The Present Study**

A careful examination of the extant psychometric literature on the SDQ however reveals a paucity of studies into whether or not scores obtained from the instrument are valid and reliable over time. Kersten et al. (2016) in their systematic review of the SDQ raised concerns about the lack of evidence of test-retest reliability, an issue related to measurement invariance over time. As far as we are aware, only one study (Croft et al., 2015) using the United Kingdom Millennium Cohort Study has so far attempted to explore factorial invariance of the SDQ across time despite its predominant use in longitudinal research. In their study, Croft and her colleagues first explored invariance for each subscale of the parent version and found an adequate fit for the conduct problem and prosocial behaviour subscales across time. However, only weak invariance was found for other subscales. With respect to the 5 factor model, weak (metric) rather than strong (scalar) longitudinal factorial invariance was achieved suggesting the possibility of a response shift for some subscales over time (Oort, 2005). Fundamentally, weak measurement invariance raises questions about the validity of comparing key parameters such as regression coefficients, means of latent variables, and composite scores across time (Millsap, 2011). As noted by Widaman et al. (2010), absence of strong longitudinal invariance raises questions about the validity of longitudinal outcomes derived from an instrument due to a higher likelihood for ambiguous or misleading conclusions about change. In other words, the evidence for invariance of the SDQ across time is not fully established and warrants further investigation.

In the current study, we address this research gap by using data from a nationally representative sample of children in Scotland to test whether or not strong longitudinal factorial invariance for the parent version of the SDQ is tenable. Our strategy involved initially testing two theoretically competing models (3 and 5-factor). This was followed by a series of nested models with increasing levels of restrictions to establish the validity and reliability of the SDQ across time. Our study, being only one of two studies to explore longitudinal invariance with a new dataset, contributes crucial knowledge on the validity of making clinical, educational, policy, and research decisions based on scores derived from the SDQ across time. It is also one of a few studies to explore invariance in models with more than two subscales.

**Methods**

**Dataset and Participants**

We used data from the Growing Up in Scotland Survey (GUS), a longitudinal survey of children born in Scotland. To ensure a nationally representative sample, participants were selected using a multi-stage stratified random sampling technique of all eligible children within a cohort year. Data were obtained annually through face-to-face interviews of the child’s main carer (mostly mothers, 95.5% of respondents). The GUS consists of multiple cohorts (1 child, and 2 birth cohorts). Further details on the sampling procedure and method of data collection are available on the GUS website (http://growingupinscotland.org.uk/) and official user guides (Bradshaw, Corbett, & Tippings, 2013).

This study uses data from 3 waves (2008/09 to 2010/11) of the first Birth Cohort survey when children were just about 4, 5 and 6 years old. The Birth cohort survey started in 2005 when the children were 10.5 months old. Subsequent waves were obtained at 22, 34.5, 46, 58, and 70 months respectively. A total of 5212 children born between June 2004 and May 2005 were recruited for the initial survey in wave one. 3375 participants who responded to all three waves of data collection (4, 5, & 6) during which SDQ data were obtained, were retained for analysis. This represents 94.2% of all eligible respondents (those who completed all previous 5 waves) and 64.7% of all Wave 1 cases. The gender distribution of the cohort sample was 51.3% male and 48.7% female. Ethnicity of the cohort children as designated in the GUS dataset was 96.5% ‘White’ and 3.5% ‘Other ethnic background’. Equivalised household income suggests that about 15%, 19%, 20%, 24% and 22% of respondents were from the bottom, second, third, fourth and top income quintiles of Scotland respectively.

**Strength and Difficulties Questionnaire (SDQ)**

The parent-report version of the SDQ (Goodman, 1997) was used at waves 4, 5 and 6 of the GUS survey corresponding to when children were aged 4, 5 and 6 years old. Participants responded to the 25 items measuring five subscales. Each subscale consists of five items. Sample items for each subscale are: conduct problems (e.g. child often has temper tantrums or hot tempers), hyperactivity-inattention (e.g., child is restless, overactive, cannot stay still for long), emotional symptoms (e.g., child has many worries, often seems worried), peer problems (e.g., child rather solitary, tends to play alone), and prosocial behaviour (e.g., child shares readily with other children). The SDQ was administered to the parents as a self-complete module during the interview process. Parents rated the cohort child on a 3-point scale (0 – Not true; 1 – Somewhat true; 2 – Certainly true).

Mean values of items on the difficulties scales ranged from 0.03 (Steals) to 1.02 (Thinks), while those on the prosocial scale ranged from 1.42 (Shares) to 1.81 (Kind). Skew and kurtosis indices of difficulties scale items ranged from -0.01 to 3.69, and -1.04 to 13.99 respectively. However, one item (Steals) had outlier skew and kurtosis values of 7.58 and 61.82. Values for items on the prosocial scale ranged from -2.09 to -0.28 and -0.24 to 3.56 respectively. Given our sample size and analysis procedure, these deviations from normality did not affect our findings. Subscale scores (from 0 – 10) and a total difficulties score (from 0 – 40) which excludes the prosocial subscale can be computed for each participant based on composite score of items. Weighted mean and standard deviation scores in this sample were low for conduct problems (*M1*=1.97, *SD1*=1.43; *M2*=1.76, *SD2*=1.44; *M3*=1.60, *SD3*=1.45), hyperactivity-inattention (*M1*=3.69, *SD1*=12.22; *M2*=3.76, *SD2*=12.34; *M3*=3.61, *SD3*=2.41), emotional symptoms (*M1*=1.19, *SD1*=1.38; *M2*=1.27, *SD2*=1.49; *M3*=1.28, *SD3*=1.59), and peer problems (*M1*=1.17, *SD1*=1.42; *M2*=1.04, *SD2*=1.37; *M3*=1.00, *SD3*=1.39) across time, and high for prosocial behaviour (*M1*=7.84, *SD1*=1.75; *M2*=8.21, *SD2*=1.65; *M3*=8.39, *SD3*=1.65). Using Goodman’s (1997) cut-off points for each subscale[[1]](#footnote-1), the proportion of children with ‘abnormal’ levels of difficulties across time were as follows: conduct problems (14.2%, 11.9%, and 10.4%), hyperactivity and inattention (11.3%, 12.3%, and 12.3%), emotional symptoms (2.8%, 4.1%, and 4.8%), peer problems (7.3%, 6.6%, and 6.6%). Proportion of normal levels of prosocial behaviour across time was 89.4%, 92.5% and 93.2% respectively. Overall, these figures are consistent with previous findings in non-clinical samples (A. Goodman & R. Goodman, 2011).

**Analytic Procedure**

**Testing Configural, Metric, and Scalar Invariance over time.**

To assess longitudinal measurement invariance of the SDQ, two competing and three nested models were specified and tested to the same dataset using a confirmatory factor analysis framework (Little, 2013; Widaman et al., 2010). Our approach follows three logical steps – configural, metric and scalar invariance testing (Figure 1). In the *configural model*, we explored whether the structure of the SDQ was invariant over time. Two competing models, that is, the original 5-factor model as well as the 3-factor model (externalising, internalising, and prosocial) were tested. Although studies testing alternative conceptualisations have consistently demonstrated superior structural validity of the original 5-factor model (Kersten et al., 2016; McCrory & Layte, 2012), the 3-factor model has been recommended for use in epidemiologic studies and low risk populations (Dickey & Blumberg, 2004; A. Goodman et al., 2010). In the configural model, only the structure of the models (i.e., 5 factors or 3 factors) and number of items (i.e., 25 items) were specified to be the same across time. To account for measurement obtained from the same participant over time, the covariances of the indicators errors (autocorrelated errors) were specified across each time point (Little, 2013; Newsom, 2015). All but two of the autocorrelated errors were subsequently found to be significant. Additionally, covariances were specified between all latent constructs at each time point. Both factor loadings and thresholds were freely estimated across each time points.

In step 2, a *metric invariance* model was tested using the best fitting model from the configural phase*.* Metric invariance was tested by specifying factor loadings (**λ**’s) for each item to be equal over time (e.g., **λ**2 at time 1= **λ**2 at time 2= **λ**2 at time 3; Coertjens, Donche, De Maeyer, Vanthournout, & Van Petegem, 2012). This is in addition to constraints in the configural model. Metric invariance tests if the relation between the items and factors specified in the SDQ are the same across time. Finally, *scalar invariance* was tested by specifying thresholds (**τ’**s) of items to be equal over time (e.g., **τ**2 time 1; threshold 1= **τ**2 time 2; threshold 1= **τ**2 time 3; threshold 1; Coertjens et al., 2012). Because our data was ordinal, thresholds, and not intercepts were specified (Davidov, Datler, Schmidt, & Schwartz, 2011; Millsap & Yun-Tein, 2004). Specifically, thresholds represent the response shift on an item (e.g., from ‘Not True’ to ‘Certainly True’), and scalar invariance tests whether the difficulty level of moving from one response to another for each SDQ item remains constant across time (Coertjens et al., 2012; Davidov et al., 2011; Newsom, 2015). All latent variable means were freely estimated and allowed to differ across time.

----------------------------------------

INSERT FIGURE 1 HERE

----------------------------------------

**Model estimation and missing data.**

The ordinal nature of the data required that a weighted least squares means-variance (WLSMV) estimation procedure was used. WLSMV is a distribution free procedure that yields more accurate test statistics, parameter estimates and standard errors in CFA modelling with ordinal level data (Newsom, 2015). Analyses were undertaken in Mplus 7.2. The GUS survey is a clustered and stratified sample with longitudinal weights to adjust for selection probability and attrition. Longitudinal weights were generated using sociodemographic characteristics associated with non-response (e.g., employment, income, and age) so as to increase representation of respondents with non-response characteristics in the sample. As a result, we used the complex model specification option in Mplus for our analysis. This specification takes into account the effects of data cluster, stratification and longitudinal weights in the computation of model fit indices and parameter estimates (Muthén & Muthén, 2012). Since the sample had negligible item-level missing data over time (average of 1.9%; 1.28%, and 1% across age 4, 5, & 6 respectively), the weighted least squares treatment of missing data approach implemented in Mplus which uses all observed data to produce parameter estimates was used. This approach maximises the likelihood that estimates based on the observed data come from the sample population (Asparouhov & Muthén, 2010).

**Model evaluation.**

Overall model fit was assessed using multiple indices (West, Taylor, & Wu, 2013). In the current study, we used the Tucker-Lewis Index (TLI) and the comparative fit index (CFI) with values greater than .90 and .95 indicative of “adequate” and “good” fit respectively (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004; West et al., 2013). Additionally, the root mean square error of approximation (RMSEA) values lower than .05 were indicative of good fit, with those ranging from .05 to .08 indicative of adequate fit (Browne & Cudeck, 1992; Hu & Bentler, 1999; West et al., 2013). Since competing and nested models were tested, significant changes in chi-square (ΔX2), changes in CFI of >0.01, as well as other model fit parameters must be examined to determine whether there was an unacceptable decrease in model fit for nested models (Little, 2013; Cheung & Rensvold, 2002). While a ΔX2 is traditionally used to evaluate nested models, simulation studies (Cheung & Rensvold, 2002; Meade, Johnson & Braddy, 2008) suggest that itperforms poorly against change in CFI in the context of invariance testing. This is because the ΔX2is too sensitive to marginal differences (see Little, 2013). Thus, we relied on a set of measures in evaluating nested models with greater emphasis on change in CFI. With respect to the evaluation of competing models, global fit indices were used. Finally, the significance of parameter estimates, and the interpretability of the obtained solutions were taken into account in judging a model’s acceptability.

**Results**

**Invariance of the SDQ over Time**

Table 2 presents the fit indices for each of the models tested. Results of the configural model indicate that the 5-factor model (CFI = .95; TLI=.94; RMSEA=.02) had a better fit to the data than the more parsimonious 3-factor model (CFI = .92; TLI=.92; RMSEA=.02). Since, the fit indices indicate a good model fit for the 5-factor model but only an adequate fit for the 3-factor model, only the 5-factor model was explored in subsequent analyses.

A metric invariance model restricting all factor loadings to be equal across time, in addition to constraints specified in the configural model, showed a good fit to the data (CFI = .95; TLI=.94 RMSEA=.01) with no evidence of a significant deterioration (ΔCFI=.003) in the 5-factor metric model. Finally, the scalar invariance model introducing additional restriction for all thresholds to be equal across time intervals resulted in an adequate model fit data (CFI =.95; TLI=.94; RMSEA=.01). The change in CFI was within the acceptable range (ΔCFI=-.002) suggesting no significant deterioration in model fit. Examination of modification indices did not show any significant form of local misfit.

--------------------------------------------

INSERT TABLE 2 HERE

---------------------------------------------

**Factor Loadings and Item Reliability**

The standardised factor loadings for the 5-factor model (Table 3) were all high. Average standardised factor loadings ranged from .57 to .62 for conduct problems, .69 to .75 for hyperactivity/inattention, .60 to .68 for emotional symptoms, .55 to .62 for peer problems, and .61 to .65 for prosocial behaviour. Items with the highest factor loadings on each subscale remained prominent across time (e.g., ‘restless’ for hyperactivity, and ‘unhappy’ for emotional problems). The items with the lowest factor loadings across time were ‘steals from home, school or elsewhere’ (steals), ‘often lies or cheats’ (lies) from the Conduct problem scale, and ‘often complains of headaches, stomach aches, or nausea (somatic) from the Emotional problems scale. The relatively high standardised factor loadings and the explained variances (*R*2) of the items (Table 3) suggest that, on the whole, items in the SDQ are reliable.

--------------------------------------------

INSERT TABLE 3 HERE

---------------------------------------------

**Association between SDQ Subscales across Time**

Examination of the cross-scale association among the five SDQ subscales (Table 4) indicate stronger associations between externalising dimensions (conduct and hyperactivity, *average r=.72*) and internalising dimensions (emotional and peer problems, *average r=.68*) than with other subscales. However, values were within the threshold indicative of distinctive 5 factors. Associations between emotional symptoms and prosocial behaviour were small across time (*r*=-.18 to -.22). Additionally, high correlations between the same latent factors across time (Table 4) indicate a high degree of stability of individual differences over time. Estimated latent means over time suggest a decreasing trend for conduct and peer problems, an increasing trend in emotional symptoms and prosocial behaviour, and a non-linear trend in hyperactivity. However, the significance and nature of these trends are best determined using latent growth curve analysis (Little, 2013).

--------------------------------------------

INSERT TABLE 4 HERE

---------------------------------------------

**Discussion and Implications**

According to Little (2013), factorial invariance is one of the most important empirical questions in longitudinal research. This study explored whether the SDQ, one of the most commonly used instruments in developmental research is invariant over time. Importantly, the SDQ is currently being used in several longitudinal studies and findings are used to inform educational, clinical and policy decisions. Our findings demonstrate strong longitudinal factorial invariance for the 5-factor parent version of the SDQ, and we did not find any type of response shift. Specifically, the stability of factor loadings, invariance of factor loadings and thresholds across time suggests an absence of reconceptualization, reprioritisation, or recalibration of responses on the SDQ (Oort, 2005; Schwartz & Sprangers, 1999). In other words, respondents’ understanding of items and constructs measured by the SDQ remain the same across time. It can therefore be concluded that any observed changes in SDQ scores over time is attributable to ‘true’ change in children’s behaviours and emotions, based on parental behavioural ratings (Oort, 2005; Widaman et al., 2010). The findings therefore support the use of the SDQ for longitudinal studies enabling researchers, educational personnel, clinicians and policy makers to draw stronger conclusions about developmental changes (Widaman et al., 2010). Further, the use of composite scores for subsequent analyses and decision making will be justified (Steinmetz et al., 2013).

While Croft and colleagues (2015) found weak invariance, our study confirms strong longitudinal invariance of the SDQ across time. However, the difference in findings between the two studies may be due to differences in study conditions. Firstly, in Croft et al.’s study, measurement of the SDQ were obtained when children were aged 3, 5 and 7 whereas our study was based on data obtained when children were aged 4, 5 and 6. Secondly, the version of the SDQ used at age 3 in Croft et al.’s study is a slightly modified version of the standard SDQ (4-16 years). Specifically, three items from the standard version (two conduct problems items, and one hyperactivity item) have been modified for children who are three years of age, and these items were freely estimated in Croft et al.’s study when testing for invariance. In other words, similar rather than identical instruments were used across time in their study. Our data on the other hand was based on only the standard version of the SDQ and all items were specified to be equal across time when testing for invariance.

Overall, R. Goodman’s (1997) 5-factor model was a superior fit to the data compared to the alternative 3-factor model tested. However, the 3-factor model in our study was an adequate fit at a configural level. While some previous studies have equally found an adequate fit for the 3-factor model (e.g., A. Goodman et al., 2010), this support is by no means universal (e.g., Croft et al., 2015; McCrory & Layte, 2012). Care is therefore needed in using the 3-factor model as the stability of this conceptualisation may be sample dependent. A sample specific investigation may therefore be warranted if the 3-factor model is to be employed for a longitudinal study. Closely linked to the above is the high association between conduct and hyperactivity subscales, and emotional and peer problems subscales. There is considerable conceptual overlap between these subscales which suggests the potential utility of an externalising and internalising subscale on the basis of parsimony. It also suggests that where there is pressure on respondent time, it may be possible to use only one subscale from the externalising and internalising dimensions. However, the above decisions would have to be balanced against the fact that the levels of associations between these subscales are within the benchmark indicative of distinctive 5 factors (Kline, 2005). As suggested in the literature, a 3-factor model may be useful in low risk populations and epidemiologic studies, while a 5-factor model may be of more value in high-risk samples (A. Goodman et al., 2010). Researchers who are interested in only the difficulties subscales might also consider using a 2 factor model (internalising and externalising) without the prosocial scale. A follow-up analysis of this alternative specifications suggests a good model fit at the configural level (CFI=.94; TLI=.94).

Consistent with previous findings (McCrory & Layte, 2012; Stone et al., 2015), the items with the lowest factor loadings across time were ‘steals from home, school or elsewhere’ (steals), ‘often lies or cheats’ (lies) from the Conduct problem scale, and ‘often complains of headaches, stomach aches, or nausea (somatic) from the Emotional symptom scale. Careful examination revealed that these items have limited variance with less than 0.6%, 1.4% and 2.8% of parents endorsing them (see also McCrory & Layte, 2012), and one of these items (steals) had outlier skew and kurtosis values. White, Connelly, Thompson and Wilson (2013) also found that participants often report unease with these items. While some researchers have suggested eliminating these ‘problematic items’ (Chauvin & Leonova, 2015), we did not find any consistent evidence in our study to suggest that these items are a threat to invariance. It may also be the case that these are the items that legitimately differentiate between ‘abnormal’ and ‘borderline’ difficulties.

Our study is limited by the fact that we have only explored longitudinal measurement invariance for the parental version of the scale, and are therefore unable to comment on other versions. Additionally, the current study is limited by data from three time points in early childhood. Multiple data points across early childhood, middle childhood and late childhood may provide a stronger evidence of invariance. However, the current study employs data which captures important transition time between preschool and early primary school. Further, our sample consists of a general sample of children rather than clinical samples. Thus, generalisation to clinical groups is not possible. Future studies should therefore consider exploring longitudinal invariance of the SDQ across versions and different groups. Finally, as noted by Widaman et al. (2010), achievement of measurement invariance while a useful first step cannot tell us about the utility the constructs measured. Thus, further prospective studies exploring the predictive power of the SDQ would be necessary for ascertaining the usefulness of the SDQ.

The strength of our study lies in the use of nationally representative sample, accounting for the complex structure of our sample in our analysis, and application of Schwartz and Strangers’ (1999) theoretical framework of ‘response shift’ in exploring longitudinal invariance of the SDQ. On the whole, our research is only one of two studies to investigate measurement invariance of the SDQ across time. Furthermore we have taken the ordinal measurement of the items into account by using ordinal confirmatory factor analysis. Methodologically, it is one of a few studies to explore invariance in models with more than two-factors. Considering the extent to which scores from the SDQ are used to inform important educational, clinical and policy decisions around the world, our findings provide important psychometric information necessary for drawing substantive conclusions about changes in child welfare across time.



Figure 1. Specification of the longitudinal measurement model. For simplicity of illustration we used only the 1-factor model rather than the 5-factor model as in our study

Table 1. *Relationship between Measurement Invariance Parameters and Typologies of response Shift (adapted from Oort, 2005)*

|  |  |  |
| --- | --- | --- |
| Parameters in Measurement Invariance | Outcome | Type of Response Shift |
| 1. Factor loadings equal across time | Yes | *No response shift* |
|  | No | Reprioritisation |
| 1. Intercepts/thresholds equal across time | Yes | *No response shift* |
|  | No | Recalibration |
| 1. Pattern of factor loadings stable across time | Yes | *No response shift* |
|  | No | Reconceptualization |

Table 2. *Model Fit Indices for the 5 and 3-Factor Measurement Invariance Models*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model | X2 | *df* | CFI | TLI | RMSEA [95% CI] | ΔCFI | ΔWLSMV X2 (Δ*df*) |
| Configural Invariance (3 factors) | 5314.25\*\*\* | 2589 | .921 | .92 | .02 [.02, .02] | *-* | *-* |
| Configural Invariance (5 factors) | 4432.30\*\*\* | 2520 | .95 | .94 | .02 [.01, .02] | - | - |
| Metric Invariance (5 factors) | 4338.24\*\*\* | 2560 | .95 | .94 | .01 [.01, .02] | .003 | 70.19 (40)\*\*\* |
| Scalar Invariance (5 factors) | 4519.24\*\*\* | 2650 | .95 | .94 | .01 [.01, .02] | -.002 | 643.78 (90)\*\*\* |

Note: X2 – chi-square; *df* – degree of freedom; CFI – comparative fit index; TLI – Tucker-Lewis Index; RMSEA – root mean square error of approximation with 95% confidence interval; ΔCFI – change in comparative fit index; ΔWLSMV X2  – change in WLSMV chi-square using chi-square difference test; Δ*df* – change in degree of freedom

Table 3. *Standardised Factor Loadings (β) and Item Reliability of SDQ Items at Ages 4, 5, and 6*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Age 4 |  | Age 5 |  | Age 6 |  |
|  | β | *R2* | β | *R2* | β | *R2* |
| **Conduct problem** |  |  |  |  |  |  |
| Tantrums | .62 | .38 | .65 | .43 | .66 | .44 |
| Obedient\* | .62 | .39 | .64 | .41 | .69 | .47 |
| Fights | .66 | .44 | .66 | .44 | .73 | .53 |
| Lies | .44 | .19 | .48 | .23 | .54 | .30 |
| Steals | .45 | .20 | .42 | .18 | .48 | .23 |
| **Hyperactivity** |  |  |  |  |  |  |
| Restless | .80 | .64 | .82 | .68 | .80 | .64 |
| Fidgeting or squirming | .72 | .52 | .74 | .55 | .77 | .59 |
| Distracted | .79 | .63 | .81 | .66 | .81 | .65 |
| Thinks\* | .50 | .25 | .55 | .30 | .65 | .42 |
| Persistent\* | .65 | .42 | .74 | .55 | .76 | .58 |
| **Emotional problems** |  |  |  |  |  |  |
| Somatic | .44 | .20 | .45 | .20 | .52 | .27 |
| Worries | .61 | .38 | .67 | .44 | .72 | .52 |
| Unhappy | .74 | .55 | .78 | .61 | .82 | .67 |
| Clingy | .58 | .34 | .60 | .36 | .64 | .41 |
| Fears | .64 | .41 | .70 | .49 | .69 | .47 |
| **Peer problems** |  |  |  |  |  |  |
| Solitary | .61 | .37 | .56 | .32 | .59 | .35 |
| Good friend\* | .45 | .20 | .54 | .29 | .51 | .26 |
| Popular\* | .68 | .47 | .67 | .45 | .78 | .61 |
| Bullied | .46 | .21 | .61 | .37 | .58 | .34 |
| Prefer adults | .53 | .28 | .54 | .29 | .64 | .41 |
| **Prosocial** |  |  |  |  |  |  |
| Considerate | .64 | .41 | .67 | .45 | .72 | .52 |
| Shares | .62 | .39 | .62 | .38 | .67 | .45 |
| Helpful | .59 | .35 | .65 | .42 | .63 | .40 |
| Kind | .65 | .42 | .68 | .46 | .71 | .50 |
| Helps | .53 | .28 | .50 | .25 | .54 | .29 |

Note: All variables are significant at *p<.001*. Factor loadings (β) and *R2* are from configural model. \* indicates items that were reverse coded

Table 4. *Correlation between Latent SDQ Sub-scales over Time*

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| 1.Conduct *(time1)* | - |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 2.Hyperactive *(time 1)* | .74 | - |  |  |  |  |  |  |  |  |  |  |  |  |  |
| 3.Emotion *(time 1)* | .50 | .34 | - |  |  |  |  |  |  |  |  |  |  |  |  |
| 4.Peer prob *(time 1)* | .48 | .38 | .70 | - |  |  |  |  |  |  |  |  |  |  |  |
| 5.Prosocial *(time 1)* | -.65 | -.42 | -.18 | -.52 | - |  |  |  |  |  |  |  |  |  |  |
| 6.Conduct *(time 2)* | .83 | .61 | .42 | .41 | -.48 | - |  |  |  |  |  |  |  |  |  |
| 7.Hyperactive*(time 2)* | .61 | .79 | .26 | .39 | -.32 | .70 | - |  |  |  |  |  |  |  |  |
| 8.Emotion *(time 2)* | .38 | .28 | .74 | .56 | -.18 | .51 | .33 | - |  |  |  |  |  |  |  |
| 9.Peer prob *(time 2)* | .44 | .40 | .50 | .80 | -.36 | .50 | .46 | .66 | - |  |  |  |  |  |  |
| 10.Prosocial *(time 2)* | -.45 | -.37 | -.17 | -.38 | .74 | -.66 | -.45 | -.18 | -.53 | - |  |  |  |  |  |
| 11.Conduct *(time 3)* | .79 | .57 | .37 | .39 | -.43 | .84 | .62 | .39 | .41 | -.49 | - |  |  |  |  |
| 12.Hyperactive*(time 3)* | .59 | .73 | .28 | .38 | -.30 | .61 | .82 | .30 | .42 | -.39 | .73 | - |  |  |  |
| 13.Emotion *(time 3)* | .42 | .27 | .68 | .54 | -.13 | .44 | .29 | .77 | .59 | -.16 | .51 | .37 | - |  |  |
| 14.Peer prob *(time 3)* | .44 | .41 | .50 | .70 | -.29 | .47 | .47 | .53 | .83 | -.40 | .58 | .55 | .68 | - |  |
| 15.Prosocial *(time 3)* | -.45 | -.36 | -.23 | -.41 | .66 | -.56 | -.40 | -.23 | -.45 | .77 | -.67 | -.50 | -.22 | -.51 | - |

Note: All variables are significant at *p<.001*

**References**

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing*

*data.* Retrieved from http://www.statmodel.com/download/GstrucMissingRevision.pdf.

Black, C., & Martin, C. (2015). *Mental well-being among adolescents in Scotland.*

*Profile and trends*. Edinburgh, UK: Scottish Government.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary*

*psychometrics*. New York, NY: Cambridge University Press.

Bradshaw, P., Corbett, J., & Tipping S. (2013). Growing Up in Scotland sweep 6: 2010-

2011 User Guide*. UK Data Archive Study Number 5760*.

Bradshaw, P., Lewis, G., & Hughes, T. (2014). Growing Up in Scotland:

Characteristics of pre-school provision and their association with child outcomes. Edinburgh, UK: Scottish Government.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit.

*Sociological Methods & Research*, *21*, 230–258. doi:10.1177/0049124192021002005.

Chauvin, B., & Leonova, T. (2015). The Strength and Difficulties Questionnaire

revisited in a French-speaking population: Propositions of a reduced version of the parent SDQ. *Journal of Psychoeducational Assessment*, Preprint. doi: 10.1177/0734282915604074.

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). *Assessment*, *23*(1), 63–74. doi:10.1177/1073191114568301.

Coertjens, L., Donche, V., De Maeyer, S., Vanthournout, G., & Van Petegem, P.

(2012). Longitudinal measurement invariance of Likert-type learning strategy scales: Are we using the same ruler at each wave? *Journal of Psychoeducational Assessment*, *30*(6), 577–587. doi:10.1177/0734282912438844.

Croft, S., Stride, C., Maughan, B., & Rowe, R. (2015). Validity of the Strengths and

Difficulties Questionnaire in preschool-aged children. *Pediatrics*, *135*(5), 1–10. doi: https://doi.org/10.1542/peds.2014-2920.

Davidov, E., Datler, G., Schmidt, P., & Schwartz, S. H. (2011). Testing the invariance

of values in the Benelux countries with the European social survey: Accounting for ordinality. In E. Davidov, P. Schmidt, & J. Billiet (Eds.). *Cross cultural analysis: Methods and applications* (pp.149 – 172). New York, NY: Routledge.

Department for Education. (2012). *Guidance on data collection on the emotional health*

*of looked after children*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment\_data/file/253566/ssda903\_sdq\_guidance\_2012-13\_version1-0.pdf.

Department for Education. (2015). Mental health and behaviour in schools:

Departmental advice for school staff. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment\_data/file/416786/Mental\_Health\_and\_Behaviour\_-\_Information\_and\_Tools\_for\_Schools\_240515.pdf.

Dickey, W. C., & Blumberg, S. J. (2004). Revisiting the factor structure of the strengths

and difficulties questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry*, *43*(9), 1159–1167.

Fergusson, D. M., Horwood, L. J., & Ridder, E. M. (2005). Show me the child at seven:

The consequences of conduct problems in childhood for psychosocial functioning in adulthood. *Journal of Child Psychology and Psychiatry*, *46*(8), 837–849. doi:10.1111/j.1469-7610.2004.00387.x.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note.

*Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. doi:10.1111/j.1469-7610.1997.tb01545.x.

Goodman, A., & Goodman, R. (2011). Population mean scores predict child mental

disorder rates: Validating SDQ prevalence estimators in Britain. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *52*(1), 100–108.doi:10.1111/j.1469-7610.2010.02278.x.

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.

Hastings, R. P., & Bham, M. S. (2003). The relationship between student behaviour

patterns and teacher burnout. *School Psychology International*, *24*(1), 115–127. doi:10.1177/0143034303024001905.

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

analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55. doi:10.1080/10705519909540118.

Kersten, P., Czuba, K., McPherson, K., Dudley, M., Elder, H., Tauroa, R., & Vandal, A.

(2016). A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. *International Journal of Behavioral Development*, *40*(1), 64–75.doi:10.1177/0165025415570647.

Kline, R. B. (2005*). Principles and practice of structural equation modeling*. New

York, NY: Guilford.

Little, T. D. (2013). *Longitudinal structural equation modeling*. New York, NY:

Guilford.

Marsh, H. W., Hau, K., & Wen, Z. (2004). In search of golden rules: Comment on

hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler’s (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

Masten, A. S., Roisman, G. I., Long, J. D., Burt, K. B., Obradovic, J., Riley, J. R.,

…Tellegen, A. (2005). Developmental cascades: Linking academic achievement and externalising and internalising symptoms over 20 years. *Developmental Psychology*, *41*(5), 733–746. doi:10.1207/s15328007sem1103\_2.

McCrory, C., & Layte, R. (2012). Testing competing models of the Strengths and

Difficulties Questionnaire’s (SDQ’s) factor structure for the parent-informant instrument. *Personality and Individual Differences*, *52*, 882–887. doi:10.1016/j.paid.2012.02.011.

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), 568–92. doi:10.1037/0021-9010.93.3.568.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance.

*Psychometrika*, *58*, 525–543.

Mejia, T. M., & Hoglund, W. L. G. (2016). Do children's adjustment problems

contribute to teacher–child relationship quality? Support for a child-driven model. *Early Childhood Research Quarterly*, *34*, 13–26. doi:10.1016/j.ecresq.2015.08.003.

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York,

NY: Routledge.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-

categorical measures. *Multivariate Behavioral Research*, *39*(3), 479–515.doi:10.1207/S15327906MBR3903\_4.

Moffitt, T. E. (1993). Adolescence-limited and life-course-persistent antisocial

behavior. *Psychological Review*, *100*, 674–701.

Muthén, L. K, Muthén, B. O. (2012). *Mplus user’s guide (7th ed.)*. Los Angeles, LA:

Muthén & Muthén.

Neece, C. L., Green, S. A., & Baker, B. L. (2012). Parenting stress and child behavior

problems: A transactional relationship across time. *American Journal on Intellectual and Developmental Disabilities*, *117*(1), 48–66. doi:10.1352/1944-7558-117.1.48.

Newsom, J. T. (2015). *Longitudinal structural equation modeling: A comprehensive*

*introduction*. New York, NY: Routledge.

Normandeau, S., & Guay, F. (1998). Preschool behavior and first-grade school

achievement: The mediational role of cognitive self-control. *Journal of Educational Psychology*, *90*, 111–121.

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true

change. *Quality of Life Research*, *14*, 587–598.

Reynolds, S., MacKay, T., & Kearney, M. (2009). Nurture groups: A large scale,

controlled study of effects on development and academic attainment. *British Journal of Special Education*, *36*(4), 204–212. doi:10.1111/j.1467-8578.2009.00445.x.

Schwartz, C. E., & Sprangers, M. A. G. (1999). Methodological approaches for

assessing response shift in longitudinal health related quality-of-life research. *Social Science & Medicine*, *48*, 1531–1548. doi:10.1016/S0277-9536(99)00047-7.

Scott, S., Knapp, M., Henderson, J., & Maughan, B. (2001). Financial cost of social

exclusion: Follow up study of antisocial children into adulthood. *British Medical Journal*, *323*, 1–5. doi: http://dx.doi.org/10.1136/bmj.323.7306.191.

Scottish Government. (2008). *The Early Years Framework*. Edinburgh, UK: Scottish

Government.

Steinmetz, H., Frese, M., & Schmidt, P. (2013). A longitudinal panel study on

antecedents and outcomes of work–home interference. *Journal of Vocational Behavior*, *73*(2), 231–241. doi:10.1016/j.jvb.2008.05.002.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2010).

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*, *13*(3), 254–274. doi:10.1007/s10567-010-0071-2.

Van Sonsbeek, M. A., Hutschemaekers, G. G., Veerman, J. W., & Tiemens, B. B.

(2014). Effective components of feedback from Routine Outcome Monitoring (ROM) in youth mental health care: Study protocol of a three-arm parallel-group randomized controlled trial. *BMC Psychiatry*, *14*, 1–11. doi:10.1186/1471-244X-14-3.

West, S. G., Taylor, A. B. & Wu, W. (2012). Model fit and model selection in structural

equation modeling. In Rick. H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp 209–231). New York, NY: Guilford Press.

White, J., Connelly, G., Thompson, L., & Wilson, P. (2013). Assessing wellbeing at

school entry using the Strengths and Difficulties Questionnaire: Professional perspectives. *Educational Research*, *55*(1), 87–98. doi:10.1080/00131881.2013.767027.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within

longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. doi:10.1111/j.1750-8606.2009.00110.x.

Wilson, P., Bradshaw, P., Tipping, S., Henderson, M., Der, G., & Minnis, H. (2013).

What predicts persistent early conduct problems? Evidence from the growing Up in Scotland cohort. *Journal of Epidemiology Community Health*, *67*, 76–80. doi:10.1136/jech-2011-200856.

1. Cut-off points for each subscale are based on bandings derived from population surveys. The bandings were derived such that 80% of children scored ‘normal’, 10% ‘borderline’ and 10% ‘abnormal’. Specific cut-off points for ‘abnormal’ scores are as follows: Conduct (4-10); Hyperactivity (7-10); Emotional (5-10); Peer problem (4-10). Score of 6-10 represent ‘normal’ levels of Prosocial Behaviour. For further details see <http://www.sdqinfo.com/py/sdqinfo/b3.py?language=Englishqz%28UK%29> [↑](#footnote-ref-1)