Assessment
1–19
© The Author(s) 2018
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1073191118780461
journals.sagepub.com/home/asm
SAGE

Is Small Still Beautiful for the Strengths and Difficulties Questionnaire? Novel Findings Using Exploratory Structural Equation Modeling

Luis Eduardo Garrido[1,2], Juan Ramón Barrada[3], José Armando Aguasvivas[4], Agustín Martínez-Molina[5], Víctor B. Arias[6], Hudson F. Golino[1], Eva Legaz[7], Gloria Ferrís[7], and Luis Rojo-Moreno[8]

Abstract

During the present decade a large body of research has employed confirmatory factor analysis (CFA) to evaluate the factor structure of the Strengths and Difficulties Questionnaire (SDQ) across multiple languages and cultures. However, because CFA can produce strongly biased estimations when the population cross-loadings differ meaningfully from zero, it may not be the most appropriate framework to model the SDQ responses. With this in mind, the current study sought to assess the factorial structure of the SDQ using the more flexible exploratory structural equation modeling approach. Using a large-scale Spanish sample composed of 67,253 youths aged between 10 and 18 years (M = 14.16, SD = 1.07), the results showed that CFA provided a severely biased and overly optimistic assessment of the underlying structure of the SDQ. In contrast, exploratory structural equation modeling revealed a generally weak factorial structure, including questionable indicators with large cross-loadings, multiple error correlations, and significant wording variance. A subsequent Monte Carlo study showed that sample sizes greater than 4,000 would be needed to adequately recover the SDQ loading structure. The findings from this study prevent recommending the SDQ as a screening tool and suggest caution when interpreting previous results in the literature based on CFA modeling.

Keywords

ESEM, CFA, factor structure, adolescents, SDQ, behavioral problems, dimensionality

Mental health problems constitute a large proportion of the disease burden in young people, with 10% to 20% of children and adolescents worldwide suffering from a disabling mental illness (Belfer, 2008; Kieling et al., 2011; Patel, Flisher, Hetrick, & McGorry, 2007; Polanczyk, Salum, Sugaya, Caye, & Rohde, 2015). Poor mental health is strongly associated with health and development problems in youths, including lower educational achievements, substance abuse, violence, poor reproductive and sexual health, and suicide, which is the third leading cause of death among adolescents (Belfer, 2008; Cook et al., 2017; Patel et al., 2007). In general, it appears that girls have more internalizing problems (e.g., anxiety, depression, somatic complaints), while boys exhibit greater externalizing problems (e.g., rule-breaking behavior, aggressive behavior), and that older adolescents tend to report more problems than younger adolescents (Rescorla et al., 2007).

Because up to 50% of all adult mental disorders have their onset in childhood and adolescence, it is vital to understand their magnitude, risk factors, and progression in youth, in order to more effectively transition to a paradigm of prevention and early intervention (Belfer, 2008, Kieling et al., 2011; Merikangas, Nakamura, & Kessler, 2009; Merikangas et al., 2010). For this reason, it is important to develop reliable and valid screening tools that can facilitate

[1]University of Virginia, Charlottesville, VA, USA
[2]Pontificia Universidad Católica Madre y Maestra, Santo Domingo, Dominican Republic
[3]Universidad de Zaragoza, Teruel, Spain
[4]Basque Center on Cognition, Brain and Language, Donostia-San Sebastián, Spain
[5]Universidad de Zaragoza, Teruel, Spain
[6]Universidad de Salamanca, Salamanca, Spain
[7]Dirección General de Salud Pública, Consellería de Sanidad, Valencia, Spain
[8]Universitat de València, València, Spain

Corresponding Author:
Luis Eduardo Garrido, Department of Psychology, University of Virginia, 102 Gilmer Hall, 485 McCormick Road, Charlottesville, VA 22903, USA.
Email: leg4v@virginia.edu

early detection and prevention of mental health problems in childhood (Lundh, Wångby-Lundh, & Bjärehed, 2008; Polanczyk et al., 2015). Some of the most commonly used mental health screening instruments available for children and adolescents include the Child Behavior Checklist (Achenbach, 1991), the Behavioral Assessment System for Children (Reynolds & Kamphaus, 1992), and the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997), all of which provide a multi-informant approach to measuring childhood behavioral and emotional functioning.

The SDQ, which is the focus of this study, is a one-page questionnaire that assesses the psychological adjustment of children and adolescents across 25 attributes, some positive and others negative, with the possibility of being completed by parents, teachers, and youths themselves (Goodman, 2001). The instrument is composed of five scales that purportedly measure Hyperactivity/Inattention (HI), Emotional Symptoms (ES), Conduct Problems (CP), Peer Problems (PP), and Prosocial Behavior (PB; Goodman, 1997). It has been recommended and adopted as a routine screening and outcome measure in many countries, and is currently translated into over 60 languages (Caci, Morin, & Tran, 2015; Goodman, 2001; Stone, Otten, Engels, Vermulst, & Janssens, 2010; Warnick, Bracken, & Kasl, 2008). Additionally, the SDQ has proven to be especially popular among clinicians for several reasons, including its short administration time of around 5 minutes—its "small" length compared with similar instruments has been dubbed as "beautiful" (Goodman & Scott, 1999)—its cost-free nature, and because it covers key aspects of common childhood and adolescence psychopathology (Mathai, Anderson, & Bourne, 2004; Niclasen, Skovgaard, Andersen, Sømhovd, & Obel, 2013). Furthermore, the strengths and difficulties approach of the SDQ makes it more acceptable for parents, particularly to those in the general population (Niclasen et al., 2013).

Controversy Regarding the Factor Structure of the Strengths and Difficulties Questionnaire

Despite its widespread use and apparent screening sensitivity, the factor structure of the SDQ has been a subject of controversy in the literature. Goodman (1997) originally developed the five-scale instrument based on the factor analytic findings of the Rutter questionnaires (Elander & Rutter, 1996). Although Goodman did not subject the SDQ to factor analysis in his initial study, the theorized *five-factor* structure of the SDQ was reproduced in early studies via principal component analysis (e.g., Goodman, 2001; Smedje, Broman, Hetta, & von Knorring, 1999). However, in the years that have followed numerous studies have emerged questioning the suitability of this latent structure, citing either a poor fit to the data (e.g., Mellor & Stokes, 2007; Patalay, Hayes, Deighton, & Wolpert, 2016), nonemergence of the theoretical factors (e.g., Kim, Ahn, & Min, 2015; Mansbach-Kleinfeld, Apter, Farbstein, Levine, & Ponizovsky, 2010), or very low internal consistencies for some of the scales (e.g., Capron, Thérond, & Duyme, 2007; Du, Kou, & Coghill, 2008). Furthermore, various studies have provided support for alternative models of *three factors* (PB, Internalization [ES + PP items], and Externalization [HI + CP items]; Essau et al., 2012; Gómez-Beneyto et al., 2013; Ruchkin, Jones, Vermeiren, & Schwab-Stone, 2008), *four factors* (either as PB, HI, CP, and Internalization, or PB, ES, PP, and Externalization; Liu et al., 2013; van de Looij-Jansen, Goedhart, de Wilde, & Treffers, 2011), and models with a *positive wording factor* (Hoofs, Jansen, Mohren, Jansen, & Kant, 2015; McCrory & Layte, 2012; Palmieri & Smith, 2007; Van Roy, Veenstra, & Clench-Aas, 2008).

The determination of an optimal factor structure for the SDQ is especially complex due to its multicultural, multilingual, and multi-informant nature. For example, in their meta-analysis of 48 studies, Stone et al. (2010) found that the reliability of the SDQ scales was substantially higher for teachers than for parents, which may be due to the items being more one dimensional for teachers as a result of halo effects (Niclasen et al., 2013; Stone et al., 2010). Likewise, Stevanovic et al. (2015) noted that the factor structure of the self-report SDQ has been particularly difficult to replicate across different ethnic/cultural groups. Yet another issue is the reverse-coded items included in the SDQ, which tend to have a negative effect in the goodness of fit of the factor models and oftentimes produces large cross-loadings that cannot be explained by theory (Percy, McCrystal, & Higgins, 2008; van de Looij-Jansen et al., 2011).

An important issue that has perhaps not received enough attention is the appropriateness of using confirmatory factor analysis (CFA) to assess the factor structure underlying the SDQ responses. In their literature review, Caci et al. (2015) identified 53 published studies that evaluated the internal structure of the SDQ scores. Of these, 62.3% used CFA (41.5% alone and 20.8% in combination with exploratory factor analysis [EFA] or principal component analysis). Since 2010, the use of CFA has become even more prevalent, with 17 of the 21 (80.9%) published studies identified in Caci et al. (2015) using CFA to make a final determination on the factor structure underlying its scores. However, researchers have called into question the suitability of CFA to model the responses to psychological scales, which are generally composed of items that are not pure or infallible indicators of a single factor (Asparouhov & Muthén, 2009; Guay, Morin, Litalien, Valois, & Vallerand, 2014; Marsh, Morin, Parker, & Kaur, 2014). In this regard, it has been shown that fixing all or the majority of the items' cross-loadings to zero, as it is done in CFA, can produce biased estimations of the specified parameters, including substantially inflated factor correlations and distorted structural paths, if these cross-loadings are meaningfully different from zero in the population (Asparouhov & Muthén, 2009; Hsu, Skidmore, Li, & Thompson, 2014; Schmitt & Sass, 2011).

Using Exploratory Structural Equation Modeling to Assess the Factor Structure of the Strengths and Difficulties Questionnaire

Exploratory structural equation modeling (ESEM; Asparouhov & Muthén, 2009) is a modeling framework that can be seen as a generalization of EFA and CFA. Both EFA and ESEM specify *unrestricted* factor models (where variables are allowed to load on all the extracted factors) and produce the same measures of fit and factor loadings given the same estimators and rotation algorithms. Also, a priori theory can be tested for both EFA and ESEM for the factor loadings (Does the variable load significantly on its theorized factor?) using target rotation, and for the factor model (Does the specified model fit the data?), using the chi-square test of exact fit or fit indices (Asparouhov & Muthén, 2009). However, ESEM has much greater modeling flexibility because, unlike EFA, it can provide local measures of parameter fit, can accommodate correlated residuals, allows for measurement and structural invariance testing, can be incorporated into broader structural models, as well as models with method factors, covariates, and direct effects, among others (Asparouhov & Muthén, 2009; Marsh et al., 2014).

The ESEM framework can also be considered a generalization of CFAs in that the former specifies an unrestricted model where all cross-loadings are estimated, while the latter posits a *restricted* model where all or the majority of the cross-loadings are fixed to zero. Indeed, formal tests can be carried out to compare the two models—along with detailed examinations of parameter estimates—, for cases where the CFA is nested within the more general ESEM (Asparouhov & Muthén, 2009; Marsh et al., 2014). Furthermore, and notwithstanding the loss in parsimony, ESEM is able to accurately recover the factor structure of population models composed of independent clusters (where all cross-loadings are equal to zero; Morin, Arens, & Marsh, 2016). Additionally, depending on the nature of the research application, the available theory, and the data, ESEM can be used as an exploratory or a confirmatory tool (Marsh et al., 2014). Because it posits an unrestricted model, it is more amenable than CFA to exploratory data-driven studies where the available theory may be limited. However, it may also be used in a confirmatory manner similar to CFA, to test an expected factor structure on a new sample. This confirmatory application of ESEM is formalized by the use of target rotation, where the research analyst has much greater a priori control on the expected factor structure (Asparouhov & Muthén, 2009; Marsh et al., 2014).

The ESEM framework appears to be especially advantageous to assess the factorial structure of the SDQ scores. As noted previously, items from psychological scales such as the SDQ are expected to be fallible indicators of the constructs they are purported to measure, making it likely that they will have residual associations with other dimensions (Marsh et al., 2014). These residual associations can be accounted for by the unrestricted structures posited by the ESEM model. In addition, the SDQ is composed of several pairs of items that have very similar content (e.g., Item 2 "I am restless, I cannot stay still for long" and Item 10 "I am constantly fidgeting or squirming"), which have been found to produce stable correlated residuals (θ) across cultures (Bøe, Hysing, Skogen, & Breivik, 2016). ESEM, unlike traditional EFA, can be used to model these correlated residuals in the context of unrestricted factor analysis. Also, because the SDQ includes a combination of positively and negatively worded items, it is prone to generating wording method variance (Hoofs et al., 2015; McCrory & Layte, 2012; Van Roy et al., 2008), which can have a negative impact on the validity and reliability of its scores (Marsh, Scalas, & Nagengast, 2010; Weijters, Baumgartner, & Schillewaert, 2013; Woods, 2006). With ESEM, this wording method variance can be accounted for by the latent method factor strategy (Marsh et al., 2010).

The Present Study

The main objective of this study was to conduct a systematic assessment of the latent structure underlying the scores of the youth self-reported SDQ using ESEM. In this regard, we will show that a better understanding of its factor structure may be gained by taking advantage of the power and flexibility of the ESEM approach. Additionally, we will demonstrate how independent clusters CFA can distort the factorial structure of the SDQ scores in ways that can meaningfully affect decisions regarding their nature and usefulness. At the moment, we are only aware of one study (Chiorri, Hall, Casely-Hayford, & Malmberg, 2016) that has used ESEM to systematically evaluate the factor structure of the SDQ scores; however, the results provided by Chiorri et al. (2016) are difficult to interpret because their reported pattern matrices included numerous standardized item loadings greater than one (some as large as 1.88), which in most cases is a signal of model misspecification.

Another goal of the current study was to estimate the necessary sample size needed to accurately recover the structure of the SDQ scores. Because much larger sample sizes may be needed to recover unrestricted ESEM structures that are defined by only a small number of items per factor or that have items with moderate or low factor loadings (Schmitt, 2011), it is important to identify the type of samples that may be needed to conduct factorial studies of the SDQ. Indeed, it is possible that inconsistent findings in the literature may be partly due to some studies not having large enough samples to obtain accurate estimations. In order to achieve this goal, a Monte Carlo study was carried out to determine the congruence between the factor structure obtained with the full sample and those estimated at systematically varied sample sizes.

Method

Participants and Procedure

The initial sample was composed of 67,881 students attending secondary schools in the Valencian Community, Spain, during the 2003-2004, 2006-2007, and 2007-2008 academic years that provided information about gender and age. Following the recommendations of Hair, Black, Babin, and Anderson (2010), cases that had missing responses on more than 50% of the items of the SDQ were eliminated from the database; no further screening of the database was conducted after the deletion of these cases. Thus, the final sample was composed of 67,253 cases. Overall, 0.44% of the total number of responses was missing, with a minimum and maximum of 0.21% and 0.67% for individual items, respectively. Because the amount of missingness was very small (<2%), a single imputation of the missing data could be considered appropriate (Widaman, 2006). Thus, the missing values were imputed using the expectation-maximization algorithm with normally distributed errors, a superior technique for single imputation of missing data (Fox-Wasylyshyn & El-Masri, 2005). The gender distribution within the sample was almost equal, with 49.3% girls and 50.7% boys. The students, which were attending Grades 1st through 4th of Compulsory Secondary Education, had ages ranging between 10 and 18 years (M = 14.16, SD = 1.07) that were distributed as follows: 1 (0.0%) 10-year-old, 2 (0.0%) 11-year-olds, 103 (0.2%) 12-year-olds, 21,187 (31.5%) 13-year-olds, 24,398 (36.3%) 14-year-olds, 13,064 (19.4%) 15-year-olds, 6,722 (10.0%) 16-year-olds, 1,662 (2.5%) 17-year-olds, and 114 (0.2%) 18-year-olds.

In order to obtain the sample used for this study, regional health and education authorities from the Valencian Community, Spain, extended all secondary-level schools, including all public, charter, and private schools, an invitation to participate in a study of risk factors, early detection, and prevention of eating disorders (DICTA-CV Program). In total, 566 schools participated in the study, 312 from the Valencian province, 200 from Castellón, and 54 from Alicante. The sample was collected from the schools that accepted to participate in the study and only included those students for which passive informed consent had been obtained from their parents. After an initial assessment of the student's age and gender, the teachers handed out the survey's questionnaires, which were completed anonymously during school hours. The students did not receive any incentive to participate in this study. The current study was approved by the regional Department of Public Health (General Public Health Office of the Regional Valencian Government).

Measures

The Spanish self-report version of the SDQ (Goodman, 1997) developed by García et al. (2000) was used for the current study. The SDQ is composed of five subscales: Hyperactivity/Inattention (HI; sample item: "I am restless, I cannot stay still for long"), Conduct Problems (CP; sample item: "I get very angry and often lose my temper"), Emotional Symptoms (ES; sample item: "I am often unhappy, depressed or tearful"), Peer Problems (PP; sample item: "Other children or young people pick on me or bully me"), and Prosocial Behavior (PB; sample item: "I am helpful if someone is hurt, upset or feeling ill"). Each scale contains five items that are answered via a 3-point Likert-type scale (0 = *not true*; 1 = *somewhat true*; 2 = *certainly true*) producing scores that range from 0 to 10. Also, Goodman (1997) suggests that a total Difficulties score may be obtained by summing up the four problems subscales. Of the 25 items in the questionnaire, 10 are positively worded (5 on PB, 2 on HI, 1 on CP, and 2 on PP), while the remaining 15 are negatively worded. Additional information regarding the psychometric properties of the SDQ scores is provided in the introduction section.

Statistical Analyses

Analysis Steps. In order to assess the factor structure of the SDQ scores the sample was randomly split into two, a derivation sample (n = 33,627) and a cross-validation sample (n = 33,626). In the *first* step, the derivation sample was used to assess a wide range of ESEM models of the SDQ, perform dimensionality analyses, and evaluate potential wording effects and correlated residuals. In the *second* step, the optimal model resulting from these analyses was then tested using the cross-validation sample, so as to evaluate the stability of the derived ESEM factor structure and to test corresponding CFA models. In order to choose between an ESEM and a CFA model, we followed the guidelines proposed by Marsh et al. (2014): if the fit and parameter estimates (e.g., factor correlations, factor loadings) of the ESEM and corresponding CFA model did not differ substantially, the CFA model was preferred on the basis of parsimony; if they were meaningfully different, then the better fitting ESEM model was preferred. In the *third* step, measurement invariance of the optimal factorial structure was examined across gender and age. In the *fourth* step, the internal consistency reliability of the SDQ sum scale scores was assessed. Finally, in the *fifth* step, a Monte Carlo study was conducted to determine the necessary sample size needed to accurately recover the factorial structure of the SDQ scores.

Dimensionality Assessment. Two of the most accurate dimensionality methods available for ordinal variables were used to provide aid in the decision of the number of factors to retain: *parallel analysis* (Garrido, Abad, & Ponsoda, 2013, 2016; Horn, 1965) and *exploratory graph analysis* (EGA; Golino & Epskamp, 2017). Parallel analysis compares the eigenvalues of the empirical data set with those obtained from generated variables that are uncorrelated in the population; factors are retained as long as their eigenvalues are larger than those from their random counterparts. As recommended in the literature, parallel analysis was interpreted in conjunction with the scree test (Hayton, Allen, & Scarpello, 2004). EGA, on the other hand, is a technique that is part of a new area called network psychometrics (see Epskamp, Maris, Waldorp, & Borsboom, in press). In network psychometrics, undirected network models are used with psychological data in order to gain insight into the relationships between variables, their underlying structure, among others. With EGA, the number of latent factors is estimated by computing a Gaussian graphical model using regularized partial correlations (see Epskamp, Borsboom, & Fried, 2018). After the network model is estimated, the walktrap algorithm is used to identify which items belong to each dimension (Pons & Latapy, 2006).

Modeling Specifications. The SDQ items were factor-analyzed using the *categorical variable estimator* weighted least squares with mean- and variance-adjusted standard errors over polychoric correlations (Rhemtulla, Brosseau-Liard, & Savalei, 2012), and with geomin rotation for the ESEM structures. Item *wording effects*, which can be defined as a differential response style to positively and negatively worded items, were modeled using random intercept item factor analysis (RIIFA; Maydeu-Olivares & Coffman, 2006). With RIIFA, a separate wording method factor is modeled that is orthogonal to the substantive factors and in which all the items are specified to have the same unstandardized loading of 1 (assuming that the reversed items have *not* been recoded). Therefore, RIIFA uses only one degree of freedom (to estimate the variance of the method factor) and ensures that the method factor may only tap into wording variance by specifying an artifactual relationship between items of opposite wording polarity.

Fit Criteria. The *global fit* of the factor models was assessed with the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the Tucker–Lewis index (TLI). Values of RMSEA of less than .08 and .05 can be considered as indicative of reasonable and close fits to the data, respectively, while values of .90 and .95 may reflect acceptable and excellent fits to the data (Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004). It is important to note, nevertheless, that these cutoff values should be considered as rough guidelines and not be interpreted as "golden rules" (Marsh et al., 2004). *Local fit* was evaluated using the standardized expected parameter change statistic (SEPC). The SEPC informs of the expected standardized value a fixed parameter would obtain if it were to be freely estimated, and absolute values above .20 have been suggested as potentially signaling large misspecifications (Saris, Satorra, & van der Veld, 2009; Whittaker, 2012). In the current study, SEPCs with significant modification indices were freed one at a time (starting from the largest) until the rotated structure became stable. The rotated structure was considered unstable if for at least one factor the coefficient of congruence (c.c.; Tucker, 1951) between the solution with the error correlation fixed to zero and the solution with the error correlation freed was less than .95 (Lorenzo-Seva, & ten Berge, 2006).

Measurement Invariance. Analyses of *factorial invariance* across gender and age were conducted according to three sequential levels of measurement invariance (Marsh et al., 2014): (a) *configural invariance*, (b) *scalar (strong) invariance*, and (c) *residual (strict) invariance*. Measurement invariance was supported if, in comparison with the configural model, the fit of the restricted models did not decrease by more than .01 in CFI or increase by more than .015 in RMSEA (Chen, 2007). The theta parameterization was used for the invariance analyses. After measurement invariance was established, effect sizes for the differences in latent means across groups were computed using Cohen's *d* statistic. According to Cohen (1992), *d* values of 0.20, 0.50, and 0.80, can be considered as small, medium, and large effect sizes, respectively. Based on the literature of the SDQ, boys tend to score higher on CP and PP, while girls obtain higher scores on ES and PB (Bøe et al., 2016; He, Burstein, Schmitz, & Merikangas, 2013; Koskelainen, Sourander, & Vauras, 2001; Mellor, 2005; Van Roy, Grøholt, Heyerdahl, & Clench-Aas, 2006); the differences between boys and girls on HI have been inconsistent (Bøe et al., 2016; He et al., 2013; Koskelainen et al., 2001; Mellor, 2005). Regarding the SDQ scores across age, being younger has been associated with more CP and PP, while being older has been related to greater HI, PB, and ES (Koskelainen et al., 2001; van de Looij-Jansen et al., 2011; Van Roy et al., 2006; Yao et al., 2009).

Reliability Analysis. Internal consistency reliability for the summed scale scores was computed using the nonlinear structural equation modeling reliability coefficient (ρ_{NL}; Yang & Green, 2015), which is appropriate for ordinal indicators and can take into account correlated errors (Viladrich, Angulo-Brunet, & Doval, 2017; Yang & Green, 2015). In addition to ρ_{NL}, ordinal alpha (Zumbo, Gadermann, & Zeisser, 2007) was computed for comparative purposes, as this coefficient has been used in numerous SDQ studies (e.g., Björnsdotter, Enebrink, & Ghaderi, 2013; Bøe et al., 2016; Ortuño-Sierra et al., 2015; van de Looij-Jansen et al., 2011). It should be noted that ordinal reliability coefficients such as ordinal alpha or ordinal omega (Gadermann, Guhn, & Zumbo, 2012), do not measure the reliability of the *observed* scores but rather constitute estimates of the *hypothetical* reliability for latent scale scores based on the sum of the continuous variables that are thought to underlie the observed discrete scores (Chalmers, 2017; Yang & Green, 2015). In this regard, ordinal alpha and ordinal omega are of limited practical usefulness and should not be reported as measures of the reliability of a test's scores (Chalmers, 2017; Viladrich et al., 2017). Moreover, the alpha coefficient provides upwardly biased estimates of reliability in the presence of correlated residuals (Viladrich et al., 2017).

Monte Carlo Study. Random samples with replacement between 200 and 10,000 observations, in increments of 200, were extracted from the total sample and the optimal ESEM model derived from Steps 1 and 2 was estimated. Then, the c.c. was computed between all possible factor orderings of this estimated solution (factors with a negative c.c. were reverted) and the solution obtained with the total sample, and the alignment that produced the highest overall c.c. was retained. For each sample size evaluated, 1,000 random samples were extracted.

Analysis Software. Data handling and missing data imputation were computed using the IBM SPSS software version 20. All ESEM, CFA, and measurement invariance analyses were conducted using the M*plus* program version 7.4. The ρ_{NL} and ordinal alpha coefficients were computed with the R function *reliability* contained in the *psych* package (version 1.7.8; Revelle, 2017). Parallel analysis was computed using the MATLAB code developed by Garrido et al. (2013), which is included in the online supplemental materials (all supplementary materials are available in online version of the article). Likewise, the specifications for parallel analysis were in accordance with the recommendations by Garrido et al. (2013) for ordinal variables, including the use of polychoric correlations, eigenvalues derived from the full correlation matrix, the mean criterion, random permutations of the empirical data sets, and the generation of 1,000 random replicates. EGA was computed using the R package *EGA* (version 1; Golino, 2017). The Monte Carlo study was carried out in the MATLAB programming environment version R2014a with code developed by the authors.

Results

Descriptive statistics for the SDQ scores, including item means, standard deviations, skewness, thresholds, and polychoric correlations are presented in Supplemental Table 1. Of note in these results was the extremely large correlation of .72 between Items 2 "restless" and 10 "fidgety," which was substantially higher than the second largest correlation of .54 between Items 15 "distractible" and 25 "persistent." Also, some items showed very high levels of skewness, including Item 11 "friend" (2.83), Item 22 "steals" (2.61), Item 12 "fights" (2.36), and Item 17 "kind" (−2.35).

Derivation Analyses

The first phase of the exploration of the factor structure underlying the SDQ scores involved dimensionality assessments with parallel analysis (aided by the scree test) and EGA (Figure 1). Using the scores of the full item set (25 variables), both parallel analysis and EGA suggested that five factors be retained (see Supplemental Table 2 for the parallel analysis eigenvalues). However, parallel analysis used in conjunction with the scree test suggested that six factors might be retained, as the sixth empirical eigenvalue (1.010) was only slightly below the sixth generated eigenvalue (1.036), and there was a noticeable elbow in the plot starting at the seventh eigenvalue. Subsequent ESEM analyses revealed that there was a large error correlation between Items 2 and 10, which had an extremely high sample polychoric correlation (.72). Likewise, in the EGA (Figure 1), it can be seen that these two items formed a separate dimension. In all, these results indicated that Items 2 and 10 were largely redundant, and thus were averaged to create a new composite variable. Both parallel analysis and EGA were computed again with the new composite variable and this time they suggested that four factors be retained. Again, parallel analysis used in conjunction with the scree test indicated that five factors might be retained, as the fifth empirical eigenvalue (1.036) was just barely below the fifth generated eigenvalue (1.040) and there was a notable elbow in the plot starting at the sixth eigenvalue. Taken together, the dimensionality assessments suggested that four or five factors might be retained after taking into account the large error correlation between Items 2 and 10.

The second phase of the exploration of the SDQ structure included a systematic evaluation of sequential ESEM models from one to six factors, with increasing numbers of correlated errors, and with the inclusion of a RIIFA wording method factor. A summary of the results from these analyses is presented in Table 1, which includes the model fit statistics, wording factor loadings, and error correlations.

The decision to evaluate models with multiple error correlations was due to the high SEPC values (e.g., for $\theta_{02,10}$ the SEPCs ranged between 0.99 and 5.53) in the models without

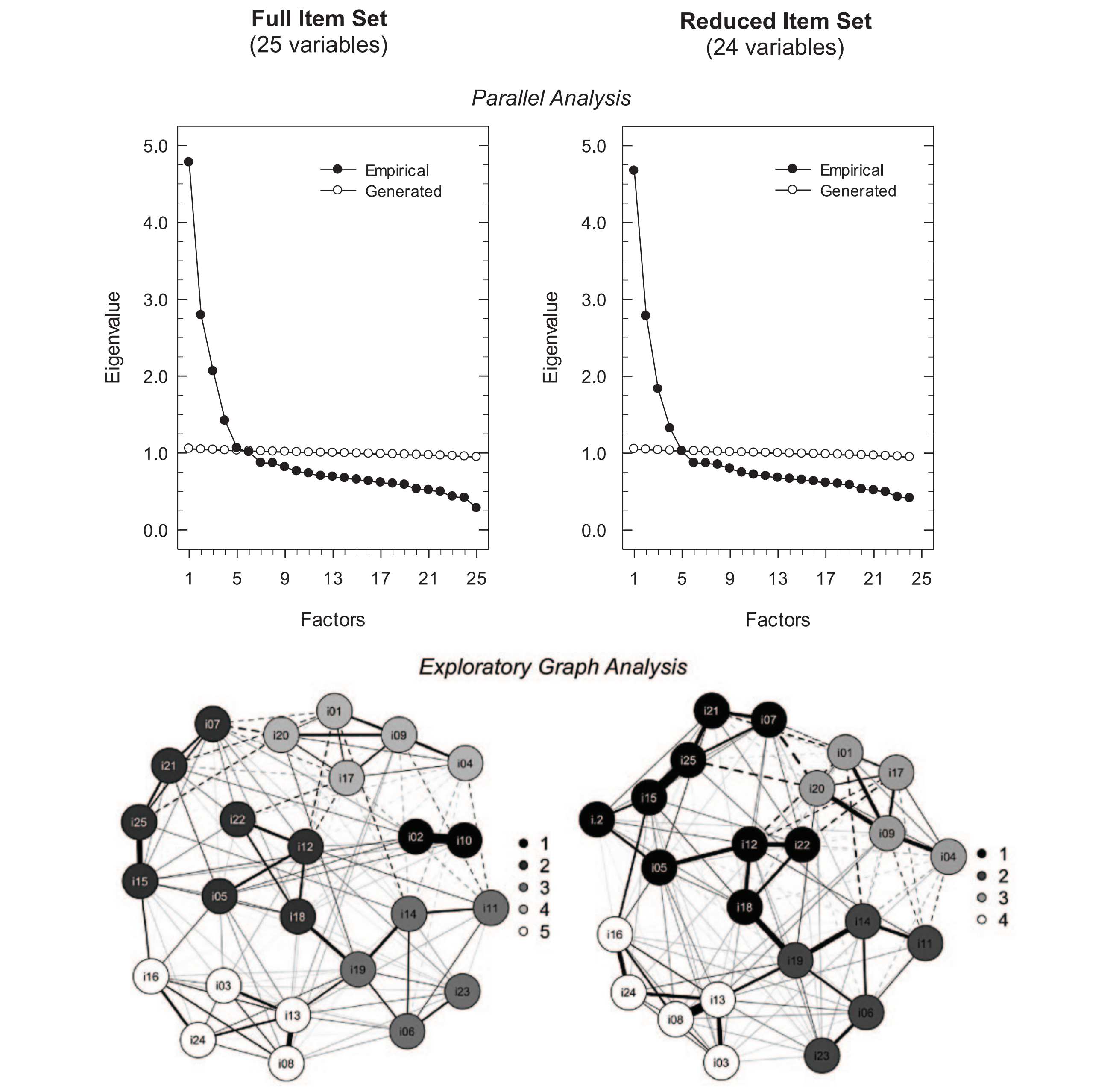

Figure 1. Dimensionality assessment using parallel analysis and exploratory graph analysis.
Note. In the reduced data set, Items 2 and 10 were averaged, decreasing the number of variables from 25 to 24 (see Item i.2 in the bottom right graph).

these error correlations (see Table 1) and because the factor structure changed meaningfully when these error correlations were estimated (see Supplemental Table 3). Specifically, for the four-factor model, two error correlations ($\theta_{02,10}$ and $\theta_{15,25}$) produced a notable change in the rotated structure when they were estimated (e.g., when adding $\theta_{02,10}$ the c.c. with the model that had zero correlated errors were .704, .998, .941, and .946, for factors one to four, respectively; when further adding $\theta_{15,25}$ the c.c. with the model that had only one correlated error were .897, .926, .922, and .468, for factors one to four, respectively), and for the five- and six-factor models, three error correlations had a discernable impact in the rotated structure when they were estimated ($\theta_{02,10}$, $\theta_{15,25}$, and $\theta_{08,13}$). In addition, an inspection of these item pairs revealed a substantial overlap in content (which appears to be even greater in the Spanish-translated version), in particular for Items 2 and 10 and Items 15 and 25 (see Supplemental Tables 10 and 11), which could help explain

Table 1. Fit Statistics, Wording Factor Loadings, and Correlated Errors for the Estimated Factor Models.

Sample/Model	χ^2	*df*	RMSEA	CFI	TLI	SEPC	WFL	θ_1	θ_2	θ_3	θ_4
Derivation sample											
ESEM-1F-0θ	63678.48	274	.083	.651	.618	$\theta_{02,10}$ = 0.99	.26				
ESEM-1F-1θ	42313.05	273	.068	.769	.746	$\theta_{15,25}$ = 0.60	.25	.67			
ESEM-1F-2θ	38004.71	272	.064	.792	.771	$\theta_{08,13}$ = 0.47	.26	.67	.51		
ESEM-1F-3θ	33971.74	271	.061	.814	.795	$\theta_{14,19}$ = 0.36	.26	.67	.50	.43	
ESEM-1F-4θ	32492.61	270	.060	.823	.803	$\theta_{06,14}$ = 0.32	.26	.67	.50	.43	.35
ESEM-2F-0θ	34668.96	250	.064	.811	.773	$\theta_{02,10}$ = 2.66	.26				
ESEM-2F-1θ	22991.69	249	.052	.875	.849	$\theta_{15,25}$ = 0.54	.20	.68			
ESEM-2F-2θ	18767.81	248	.047	.898	.877	$\theta_{08,13}$ = 0.31	.21	.67	.46		
ESEM-2F-3θ	18380.31	247	.047	.900	.879	$\theta_{14,19}$ = 0.30	.22	.67	.46	.24	
ESEM-2F-4θ	17457.07	246	.046	.905	.884	$\theta_{06,14}$ = 0.27	.22	.67	.46	.22	.28
ESEM-3F-0θ	17347.42	227	.047	.906	.875	$\theta_{02,10}$ = 3.93	.18				
ESEM-3F-1θ	9381.81	226	.035	.950	.933	$\theta_{15,25}$ = 0.64	.20	.66			
ESEM-3F-2θ	6969.52	225	.030	.963	.950	$\theta_{08,13}$ = 0.31	.21	.65	.42		
ESEM-3F-3θ	6531.93	224	.029	.965	.953	$\theta_{18,19}$ = 0.23	.21	.65	.42	.24	
ESEM-3F-4θ	6039.20	223	.028	.968	.957	$\theta_{14,19}$ = 0.24	.21	.65	.42	.24	.22
ESEM-4F-0θ	7234.60	205	.032	.961	.943	$\theta_{02,10}$ = 5.21	.19				
ESEM-4F-1θ	5634.42	204	.028	.970	.956	$\theta_{15,25}$ = 1.05	.19	.63			
ESEM-4F-2θ	4113.44	203	.024	.978	.968	$\theta_{08,13}$ = 0.31	.19	.63	.40		
ESEM-4F-3θ	3679.99	202	.023	.981	.972	$\theta_{14,19}$ = 0.22	.19	.63	.40	.24	
ESEM-4F-4θ	3387.57	201	.022	.982	.974	$\theta_{18,19}$ = 0.22	.19	.63	.39	.23	.20
ESEM-5F-0θ	4375.93	184	.026	.977	.962	$\theta_{02,10}$ = 5.53	.19				
ESEM-5F-1θ	2760.62	183	.020	.986	.977	$\theta_{15,25}$ = 0.87	.18	.63			
ESEM-5F-2θ	2447.53	182	.019	.988	.979	$\theta_{08,13}$ = 0.27	.18	.63	.34		
ESEM-5F-3θ	**2326.80**	**181**	**.019**	**.988**	**.980**	$\theta_{18,19}$ **= 0.23**	**.18**	**.64**	**.34**	**.19**	
ESEM-5F-4θ	2136.28	180	.018	.989	.982	$\theta_{14,19}$ = 0.22	.17	.63	.34	.19	.19
ESEM-6F-0θ	2414.75	164	.020	.988	.977	$\theta_{02,10}$ = 3.25	.18				
ESEM-6F-1θ	2140.18	163	.019	.989	.980	$\theta_{15,25}$ = 1.01	.18	.63			
ESEM-6F-2θ	1850.56	162	.018	.991	.983	$\theta_{08,13}$ = 0.34	.17	.63	.35		
ESEM-6F-3θ	1777.57	161	.017	.991	.983	$\theta_{18,19}$ = 0.24	.16	.63	.36	.21	
ESEM-6F-4θ	1583.06	160	.016	.992	.985	$\theta_{14,19}$ = 0.22	.16	.63	.36	.20	.19
Cross-validation sample											
ESEM-5F-3θ	**2468.60**	**181**	**.019**	**.987**	**.979**	$\theta_{18,19}$ **= 0.25**	**.18**	**.63**	**.34**	**.18**	
CFA-5F-3θ-0CL	11868.70	261	.036	.936	.926	$\lambda_{12,HI}$ = 0.50	.22	.66	.40	.26	
CFA-5F-3θ-5CL	10239.59	256	.034	.945	.935	$\lambda_{20,CP}$ = 0.93	.22	.65	.39	.25	

Note. df = degrees of freedom; RMSEA = root mean square error of approximation; CFI = comparative fit index; TLI = Tucker–Lewis index; SEPC = highest absolute standardized expected parameter change; WFL = wording factor loading; θ = error correlation; ESEM = exploratory structural equation modeling; CFA = confirmatory factor analysis; F = factors; CL = cross-loadings; λ = factor loading; HI = hyperactivity/inattention; CP = conduct problems. $p < .001$ for all chi-square tests of model fit, modification indices associated to the reported SEPCs, θs, and WFLs. $\theta_1 = \theta_{02,10}$; $\theta_2 = \theta_{15,25}$; $\theta_3 = \theta_{03,08}$; $\theta_4 = \theta_{14,19}$ (1, 2, or 4 factors) or $\theta_{18,19}$ (3, 5, or 6 factors). Values for the optimal model are bolded and highlighted in gray.

these large error correlations. Regarding the wording method factor loadings, they were significant and of notable magnitude (.16 to .26) for all ESEM models evaluated with the derivation sample. The rotated solutions for the ESEM models with one to six factors and zero to four correlated errors appear in Supplemental Tables 4 to 8.

The results shown in Table 1 reveal that the ESEM models with three or more factors and three correlated errors ($\theta_{02,10}$, $\theta_{15,25}$, and $\theta_{08,13}$) had a good fit to the data according to the conventional cutoff values (CFI, TLI ⩾ .95; RMSEA < .05). The rotated solutions for these ESEM models are shown in Table 2. As can be seen in Table 2, only one variable had a salient loading (⩾.30) in the last factor of the six-factor solution, suggesting that possibly too many factors had been extracted, which would be in line with the parallel analysis and EGA dimensionality results. In the case of the three-factor solution, the items from the PP factor had very similar salient loadings in both the second and third

Table 2. Factor Solutions for Three-, Four-, Five-, and Six-Factor ESEM Models With the Derivation Sample.

		ESEM-3F-3θ			ESEM-4F-3θ				ESEM-5F-3θ					ESEM-6F-3θ					
D	Item/factor	F1	F2	F3	F1	F2	F3	F4	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F6
HI	i02: restless	**.43**	−.01	.13	**.48**	−.01	−.04	.08	**.30**	.26	.01	−.01	.17	.17	**.35**	−.01	−.01	.18	.08
	i10: fidgety	**.37**	.02	.14	**.47**	−.03	−.03	.16	.21	**.32**	.04	−.02	.22	.05	**.41**	.02	−.03	.20	.10
	i15: distractible	**.50**	.10	.09	**.44**	.20	−.05	−.09	**.53**	.06	.08	.05	.04	**.36**	.19	.04	.04	.08	.21
	i21: reflective[a,b]	**.59**	.06	.08	**.53**	.11	.00	−.09	**.47**	.20	.02	.01	−.01	**.41**	.27	.00	.00	.02	.06
	i25: persistent[a,b]	**.46**	.13	.00	**.35**	.24	.00	−.20	**.53**	−.02	.07	.05	−.13	**.52**	.04	.06	.04	−.06	.04
CP	i05: tempers	**.42**	.16	−.01	**.42**	.13	.10	.05	.14	**.41**	.22	−.05	.04	.04	**.44**	.21	−.05	.00	.06
	i07: obedient[a,b]	**.59**	−.01	−.02	**.50**	.05	.06	−.17	**.44**	.22	−.04	−.01	−.11	**.47**	.26	−.03	−.01	−.06	−.08
	i12: fights	**.44**	.09	**−.31**	**.40**	−.03	**.41**	.01	−.07	**.64**	.10	.03	−.15	−.07	**.62**	.14	.02	−.21	−.14
	i18: lies	**.35**	.21	−.28	**.35**	.04	**.40**	.11	.13	**.36**	−.05	**.34**	−.01	.05	**.41**	−.04	**.33**	−.03	.04
	i22: steals	**.42**	−.01	−.24	**.35**	−.04	.28	−.08	.11	**.39**	−.02	.04	−.15	.07	**.40**	.00	.03	−.17	−.02
ES	i03: somatic	.09	**.43**	.04	.06	**.46**	−.02	.04	.03	.06	**.48**	−.01	.00	.11	.03	**.50**	−.01	.03	−.07
	i08: worries	.04	**.51**	.01	.04	**.50**	.03	.11	−.05	.10	**.57**	.02	.03	−.07	.09	**.54**	.03	.01	.11
	i13: unhappy	.05	**.64**	−.01	.04	**.61**	.07	.12	.03	.04	**.60**	.13	.03	.09	.02	**.61**	.14	.06	−.01
	i16: clingy	.17	**.43**	.04	.09	**.54**	−.06	−.07	.19	−.04	**.48**	.00	−.05	.01	.01	**.44**	−.02	−.08	**.42**
	i24: fears	.01	**.49**	.03	−.05	**.56**	−.04	.00	.09	−.12	**.48**	.07	−.02	.02	−.10	**.44**	.08	−.01	.24
PP	i06: solitary	−.16	**.42**	**−.45**	−.25	**.30**	**.43**	−.01	−.15	−.04	.16	**.38**	−.23	−.07	−.09	.17	**.38**	−.23	−.01
	i11: friend[a,b]	−.01	**.33**	**−.39**	−.09	.20	**.40**	−.04	−.12	.10	.12	.24	−.26	−.02	.03	.13	.25	−.28	−.05
	i14: popular[a,b]	.01	**.46**	**−.43**	−.08	**.31**	**.45**	−.02	.08	−.07	.04	**.50**	−.23	.17	−.10	.04	**.50**	−.21	.00
	i19: bullied	.00	**.50**	**−.33**	.04	.26	**.47**	.25	.04	.07	.01	**.66**	.08	.01	.10	.01	**.66**	.08	.05
	i23: adults	.01	**.30**	−.22	.02	.16	.28	.13	−.05	.11	.10	.28	.00	−.03	.11	.11	.28	−.01	−.03
PB	i01: considerate[a]	**−.37**	.02	**.35**	−.24	.06	**−.36**	.20	−.13	−.26	.13	−.11	**.32**	−.16	−.25	.12	−.11	**.33**	.03
	i04: shares[a]	−.17	−.01	**.40**	−.03	.02	**−.36**	.22	.04	−.18	.05	−.07	**.37**	.07	−.16	.07	−.07	**.42**	−.11
	i09: caring[a]	−.24	.14	**.49**	.01	.06	**−.37**	**.45**	−.02	−.12	.12	.01	**.58**	−.05	−.08	.13	.02	**.60**	−.07
	i17: kind[a]	**−.35**	.01	**.35**	−.21	.02	**−.34**	.22	−.02	**−.34**	−.01	.03	**.37**	−.05	−.30	−.02	.03	**.40**	.03
	i20: helps[a]	**−.41**	.18	.18	−.12	−.09	−.01	**.62**	**−.31**	.01	−.02	.28	**.54**	**−.38**	.03	−.02	.27	**.48**	−.02
	F1	1.0			1.0				1.0					1.0					
	F2	.26	1.0		.29	1.0			.41	1.0				.44	1.0				
	F3	−.13	.13	1.0	.21	.20	1.0		.24	.13	1.0			.19	.17	1.0			
	F4				−.22	.15	−.17	1.0	.12	.25	.47	1.0		.09	.21	.45	1.0		
	F5								−.16	−.26	.03	−.25	1.0	−.17	−.22	.06	−.24	1.0	
	F6													.32	.17	.13	.08	.07	1.0

Note. ESEM = exploratory structural equation modeling; F = factor; θ = correlated error; D = theoretical dimension; HI = Hyperactivity/Inattention; CP = Conduct Problems; ES = Emotional Symptoms; PP = Peer Problems; PB = Prosocial Behavior. Factor loadings ≥.30 in absolute value are bolded and highlighted in gray. p < .05 for all factor loadings and factor correlations, except those underlined. All models include the three correlated errors $\theta_{02,10}$, $\theta_{15,25}$, and $\theta_{08,13}$, and a wording method factor (estimates shown in Table 1).
[a]Positive polarity item. [b]Reversed item recoded.

factors, while the PB items loaded saliently in the first and third factors; these type of solutions usually signal that too few factors have been extracted, which again would be in line with parallel analysis and EGA which suggested that at least four factors be retained.

Between the four- and five-factor solutions, the one with five factors appears to be the most interpretable. In the five-factor model 22 of the 25 items had their highest loading in their theoretical factor (all except Items 7, 10, and 11), providing support for this solution. Nevertheless, it should be noted that this solution was not robust, as there were many items with weak primary loadings (<.40) and multiple salient cross-loadings. In the case of the four-factor model, the HI and CP items created an "Externalization" factor, but the items from PP and PB presented a complex structure; the last factor was composed of only two PB items, while four PB items loaded saliently in the PP factor. Also, the PP items showed some high cross-loadings in the ES factor. Thus, it appears that the five-factor model is somewhat superior to the four-factor model in terms of interpretability, a result supported by the dimensionality assessment of parallel analysis when used in conjunction with the scree test. Nevertheless, it should be noted that the decision between the four- and five-factor solutions is not a very clear one, as evidenced by the small difference in fit between the two models (.023 vs. .019 for RMSEA; .981 vs. 988 for CFI; and .972 vs. .980 for TLI) and the weak fifth eigenvalue in the parallel analysis dimensionality assessment.

Cross-Validation Analyses

The optimal five-factor ESEM structure with a wording RIIFA factor and three correlated errors ($\theta_{02,10}$, $\theta_{15,25}$, and $\theta_{08,13}$) from the derivation analyses (ESEM-5F-3θ) was tested with the cross-validation sample. In addition, two CFA models with a wording RIIFA factor and three correlated errors were evaluated: the theoretical five-factor model (CFA-5F-3θ-0CL) and a CFA model that included the five nontheoretical salient loadings (CFA-5F-3θ-5CL) observed in the ESEM derivation analyses. A summary of the results from these models is shown in Table 1 and the factor loadings and factor correlations are presented in Table 3.

As can be seen in Table 1, even though the fit of the two CFA models approximated the standard cutoff values, it was noticeably worse than the fit of the corresponding ESEM model (.036/.034 vs. .019 for RMSEA; .936/.945 vs. .987 for CFI; and .926/.935 vs. .979 for TLI). Indeed, the majority of the cross-loadings that were fixed to zero in the CFAs were significantly different from zero in the ESEM, and many had nontrivial absolute values (>.10). Additionally, when looking at the SEPCs reported in Table 1, the CFA models obtained SEPCs (.50 and .93, both for cross-loadings) that were markedly higher than the largest SEPC for the ESEM model (.25); this indicates that the CFA models displayed notable levels of local misfit that would make their acceptance questionable.

The bias introduced by fixing the cross-loadings to zero in the CFAs was evident in the estimated factor loadings and factor correlations of these models (Table 3). For example, whereas the correlation between the HI and CP factors was .37 in the ESEM, it was .87 and .81 in the CFAs. Likewise, the factor correlation between CP and PB was −.27 in the ESEM but −.75 and −.69 in the CFAs. Moreover, some of the most questionable items in the ESEM solution appeared as strong items in the CFAs, a biased result due to their unmodeled cross-loadings. For example, even though Item 7 had a .22 loading in its theoretical ESEM factor, it obtained a .55 loading in the corresponding CFA without cross-loadings. A similar result can be seen for Items 11, 18, 17, and so on. These biases were somewhat mitigated when some cross-loadings were included in the CFA-5F-3θ-5CL model, but the solution still produced a biased perspective of the quality of the SDQ items and extremely large factor correlations. Taken together, these results strongly suggest that ESEM is a more appropriate framework to model the SDQ responses than CFA. In terms of the cross-validation of the ESEM structure, all the factors obtained coefficients of congruence above .99 when the solutions from the derivation and cross-validation samples were compared, and the estimated error correlations and wording factor loadings were practically identical (see Table 1), providing strong support for the stability of this solution.

Measurement Invariance Analyses

After determining the optimal factorial structure for the total child and adolescent sample of the SDQ, the measurement invariance of this structure was evaluated across gender and age (see Table 4). In terms of age, the sample was divided between early adolescents (10-14 years) and late adolescents (15-18 years; Gore et al., 2011). As the results from Table 4 indicate, the five-factor ESEM model with a wording RIIFA factor and three correlated errors produced practically the same fit for the gender (girls/boys) and age (10-14/15-18 years) groups. Additionally, there was support for both scalar (strong) and residual (strict) levels of measurement invariance for gender and age, as the decrease in CFI in comparison with the configural model was less than .01, and the increase in RMSEA was less than .015 (in fact, the RMSEA improved [was lower] for the residual model across age groups in comparison with the configural model). In all, these results suggest that the SDQ scores had the same underlying structure and measurement properties for girls and boys, and early and late adolescents.

In order to achieve model identification for the latent mean comparisons, the means and standard deviations of the SDQ factors were fixed to 0 and 1, respectively, for the girl and early adolescent groups. When comparing the latent means across gender, boys had higher means in HI ($M = 0.066$, $p = .021$, $d = 0.065$) and CP ($M = 0.336$, $p < .001$, $d = 0.341$), lower means in ES ($M = -1.292$, $p < .001$, $d = 1.256$) and PP ($M = -0.691$, $p < .001$, $d = 0.677$), and there were no differences in PB ($M = 0.413$, $p = .329$, $d = 0.413$). Of note in these results were the large and medium effects obtained for the ES and PP factors, where girls reported substantially more problems than boys. In terms of age, late adolescents had higher means in HI ($M = 0.072$, $p = .013$, $d = 0.073$), ES ($M = 0.112$, $p < .001$, $d = 0.100$), PP ($M = 0.372$, $p < .001$, $d = 0.402$), and PB ($M = 0.872$, $p < .001$, $d = 0.807$), but a lower latent mean in CP ($M = -0.121$, $p < .001$, $d = 0.112$). In this case, the difference in PB was the greatest one, achieving a large effect size (all the other differences could be categorized as "small").

Table 3. Factor Solutions for the Five-Factor ESEM and CFA Models With the Cross-Validation Sample.

		ESEM-5F-3θ					CFA-5F-3θ-0CL					CFA-5F-3θ-5CL				
D	Item/factor	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5	F1	F2	F3	F4	F5
HI	i02: restless	**.31**	.27	−.01	.02	.17	**.36**					**.37**				
	i10: fidgety	.24	**.30**	.03	−.01	.19	**.33**					**.41**	−.08			
	i15: distractible	**.56**	.04	.06	.06	.03	**.53**					**.54**				
	i21: reflective[a,b]	**.45**	.20	.03	−.01	−.07	**.61**					**.61**				
	i25: persistent[a,b]	**.50**	−.01	.09	.03	−.15	**.57**					**.57**				
CP	i05: tempers	.16	**.39**	.23	−.04	.03		**.47**					**.49**			
	i07: obedient[a,b]	**.41**	.22	−.02	−.02	−.14		**.55**				**.38**	.22			
	i12: fights	−.06	**.62**	.11	.03	−.17		**.57**					**.60**			
	i18: lies	.15	**.33**	−.05	**.38**	.00		**.54**					**.42**		.21	
	i22: steals	.15	**.37**	−.02	.03	−.15		**.49**					**.51**			
ES	i03: somatic	.04	.05	**.48**	−.03	.00			**.46**					**.46**		
	i08: worries	−.04	.09	**.59**	.02	.03			**.52**					**.52**		
	i13: unhappy	.02	.04	**.65**	.09	.01			**.67**					**.67**		
	i16: clingy	.23	−.05	**.44**	.02	−.04			**.54**					**.54**		
	i24: fears	.12	−.14	**.45**	.08	−.02			**.48**					**.48**		
PP	i06: solitary	−.16	−.05	.11	**.40**	−.25				**.45**					**.46**	
	i11: friend[a,b]	−.16	.09	.13	.19	−.30				**.44**					**.45**	
	i14: popular[a,b]	.04	−.07	.05	**.46**	−.28				**.65**					**.65**	
	i19: bullied	.06	.03	.03	**.65**	.03				**.62**					**.63**	
	i23: adults	−.05	.12	.08	.29	.01				**.35**					**.35**	
PB	i01: considerate[a]	−.11	−.23	.11	−.09	**.37**					**.60**					**.65**
	i04: shares[a]	.01	−.13	.12	−.13	**.35**					**.42**					**.45**
	i09: caring[a]	−.02	−.09	.16	.01	**.58**					**.49**					**.53**
	i17: kind[a]	.01	**−.34**	.00	.02	**.39**					**.59**		−.25			**.33**
	i20: helps[a]	−.26	.02	−.02	.30	**.56**					**.45**	−.15				**.33**
	F1	1.0					1.0					1.0				
	F2	.37	1.0				.87	1.0				.81	1.0			
	F3	.23	.10	1.0			.47	.42	1.0			.44	.38	1.0		
	F4	.10	.23	.49	1.0		.26	.51	.60	1.0		.24	.47	.59	1.0	
	F5	−.21	−.27	−.04	−.27	1.0	−.50	−.75	−.08	−.42	1.0	−.42	−.69	−.03	−.42	1.0

Note. ESEM = exploratory structural equation modeling; F = factor; θ = correlated error; CL = cross-loading; D = theoretical dimension; HI = Hyperactivity/Inattention; CP = Conduct Problems; ES = Emotional Symptoms; PP = Peer Problems; PB = Prosocial Behavior. Factor loadings ≥.30 in absolute value are bolded and highlighted in gray. $p < .05$ for all factor loadings and factor correlations, except those underlined. All models include the three correlated errors $\theta_{02,10}$, $\theta_{15,25}$, and $\theta_{08,13}$, and a wording method factor (estimates shown in Table 1).

[a]Positive polarity item. [b]Reversed item recoded.

Internal Consistency Reliability Analyses

Because numerous studies of the SDQ have used the ordinal alpha coefficient to assess the reliability of its scale scores, we computed this coefficient for comparative purposes even though it would not be appropriate to interpret its values as estimates of observed reliability (see the Method section). Using the theoretical scales, ordinal alpha provided estimates of .69, .67, .73, .62, and .69 for the HI, CP, ES, PP, and PB scales, respectively. In contrast, and taking into account the three correlated errors included in the optimal ESEM Model ($\theta_{02,10}$, $\theta_{15,25}$, and $\theta_{08,13}$), the ρ_{NL} coefficient provided reliability estimates of .45, .53, .62, .50, and .57 for these same scales. As can be seen, the estimates of internal consistency reliability provided by ρ_{NL} are much lower than those of ordinal alpha, and the decisions regarding the reliability of the SDQ

Table 4. Measurement Invariance Analyses Across Gender and Age.

Variable	Overall model fit					Change in model fit				
Invariance model	χ^2	*df*	CFI	TLI	RMSEA	$\Delta\chi^2$	Δdf	ΔCFI	ΔTLI	ΔRMSEA
Gender										
Girls (*n* = 33,127)	2,438.8	181	.988	.979	.019					
Boys (*n* = 34,126)	2,332.1	181	.988	.980	.019					
MI1. Configural (none)	4,853.6	362	.988	.979	.019					
MI2. Scalar (FL, Th)	6,955.0	481	.982	.978	.020	2,202.0	119	−.006	−.001	.001
MI3. Residual (FL, Th, Uniq)	7,603.8	506	.980	.977	.020	2,793.7	144	−.008	−.002	.001
Age, years										
10-14 (*n* = 45,691)	3,019.0	181	.989	.981	.019					
15-18 (*n* = 21,562)	1,621.7	181	.987	.979	.019					
MI1. Configural (none)	4,705.0	362	.988	.980	.019					
MI2. Scalar (FL, Th)	6,359.5	481	.984	.980	.019	1,892.0	119	−.004	.000	.000
MI3. Residual (FL, Th, Uniq)	6,169.3	506	.984	.982	.018	1,908.4	144	−.004	.002	−.001

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; MI = measurement invariance; FL = factor loadings; Th = thresholds; Uniq = uniquenesses. The parameters constrained to be equal across groups are shown in the parentheses next to the invariance models. The chi-square difference tests between nested models was conducted using M*plus*' DIFFTEST option. *p* < .001 for all chi-square tests.

scale scores would differ if the latter were to be used. As it stands, ρ_{NL} shows that none of the SDQ scale scores approximate the minimum levels of reliability recommended for diagnostic or screening purposes (⩾.70).

Monte Carlo Study

The final assessment of the factorial structure of the SDQ scores involved the determination of the sample size needed to obtain an accurate recovery of the optimal ESEM structure obtained from the previous derivation and cross-validation analyses. The box plots in Figure 2 depict the coefficients of congruence between the ESEM solutions at sample sizes ranging between 200 and 10,000 and the estimated structure with the total sample. These results show that very large sample sizes of approximately 4,200 observations would be required to achieve a mean c.c. of .950. Moreover, even for samples this large, more than 25% of the estimated solutions still obtained coefficients of congruence lower than .950. Indeed, to have at least 90% of the solutions achieve a c.c. of .950 or greater, sample sizes of 7,000 or more observations would be needed. Additional results presented in Supplemental Table 9 and Supplemental Figures 1 to 5 show that some SDQ factors are more robust than others. For example, the CP and ES factors achieved a mean c.c. of .950 with sample sizes of 3,000 or greater, whereas the PB factor needed samples of at least 7,400 observations to achieve this same level of factor congruence. Taken together, these results indicate that typical sample sizes used in factor analytic studies (≤1,000) would not be nearly enough to obtain an accurate recovery of the population structure of the SDQ scores.

Discussion

The assessment of factorial validity is an integral component to the determination of how well instruments are able to measure underlying theoretical constructs, often dictating their potential usefulness for quantitative research, clinical diagnosis or screening, and theory development. Although the literature addressing the psychometric properties of the SDQ scores is substantial (Bøe et al., 2016; Stone et al., 2010), its interpretability may not be straightforward. This is because the techniques that have been used to factor analyze the SDQ, primarily CFA (but also EFA), appear to be unequipped to model the complex psychological mechanisms that account for the variability of its scores, possibly leading to biased results and suboptimal decisions. As a result, we sought to gain a greater understanding of the factorial validity of the SDQ scores by using the more flexible ESEM framework to conduct a systematic assessment of the latent structure underlying the scores from a large-scale Spanish adolescent sample. The main findings from this study are summarized next.

Main Findings

The results from the derivation and cross-validation ESEM analyses showed some support for the five-factor theoretical structure proposed by Goodman (1997), with the five-factor model providing an interpretable solution that had a good fit to the SDQ responses and that was invariant across gender and age. However, it also shed light into several problematic issues such as the presence of a number of questionable indicators, multiple residual correlations, wording effects, and a generally

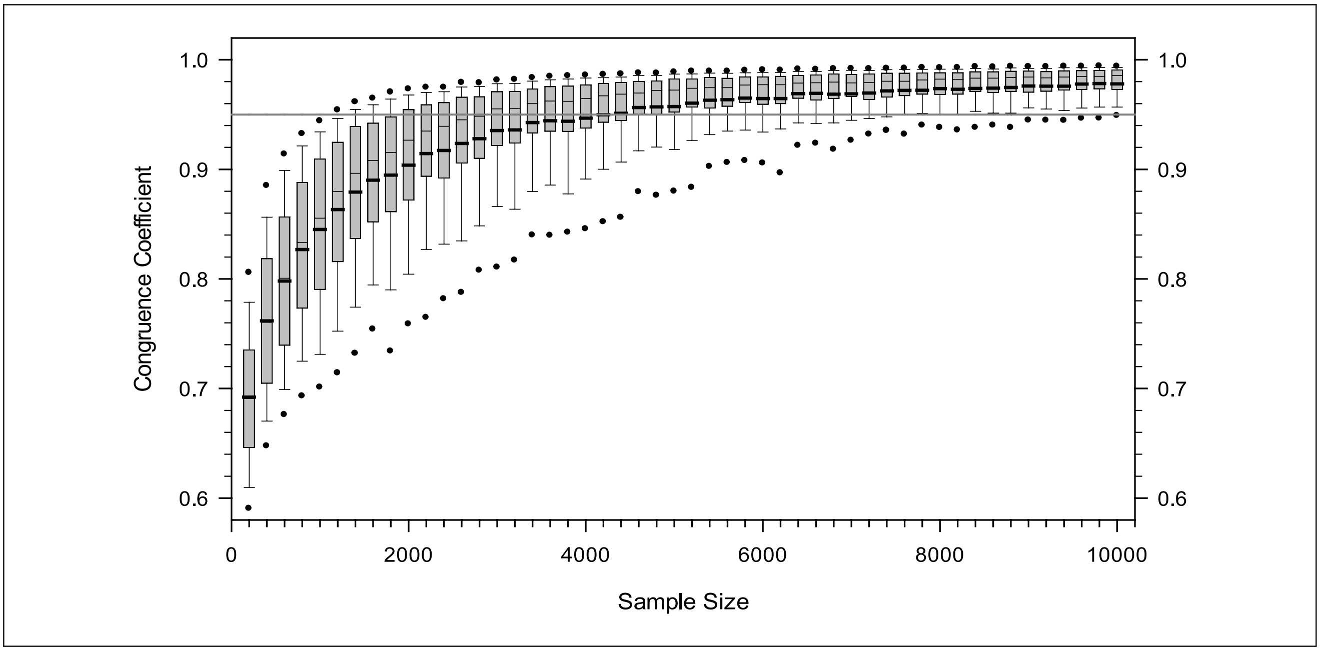

Figure 2. Box plots representing the factor loading congruence between the optimal five-factor ESEM solution with the total sample (*N* = 67,253) and the solutions obtained with the extracted random samples.

Note. The thick horizontal lines within each box represent the mean coefficient of congruence for each sample size; the thin horizontal lines represent the median values. The top and bottom black circles indicate the 95th and 5th percentiles, respectively. The horizontal gray line represents a coefficient of congruence of .950.

weak and unstable factorial structure. Complementary dimensionality assessments using parallel analysis and EGA also provided partial support for a five-factor SDQ structure, suggesting either a five-factor solution with a very weak fifth factor or a four-factor structure.

Overall, 7 of the 25 SDQ items (28%) were identified as questionable indicators of their theoretical constructs through the ESEM analyses. Four of these items had cross-loadings that were higher in absolute magnitude than their theoretical loading: Items 7 ("obedient," CP), 10 ("fidgety," HI), 11 ("friend," PP), and 18 ("lies," CP). Also, Items 11 and 23 ("adults," PP) did not achieve a salient loading (⩾.30) on their theoretical factor, and Items 2 ("restless," HI) and 17 ("kind," PB) had main loadings that were less than .05 above their highest cross-loading. In addition to these questionable indicators, three error correlations were identified through the ESEM analyses that were of notable magnitude (⩾.20; Whittaker, 2012) and/or that had a discernible impact on the rotated structure when the parameter was freed. The two largest of these error correlations included item pairs from the HI dimension: Items 2 "restless" and 10 "fidgety," and Items 15 "distractible" and 25 "persistent." Error correlations between these item pairs have been reported in a great number of SDQ studies across many cultures and languages (e.g., Bøe et al., 2016; Niclasen et al., 2013; Ortuño-Sierra et al., 2015; Percy et al., 2008; Tobia, Gabriele, & Marzocchi, 2013; van de Looij-Jansen et al., 2011; Van Roy et al., 2008).

Although similar wording might help explain the positive error correlations between Items 2 and 10 (which refer to hyperactivity) and Items 15 and 25 (which refer to attention deficit), they could also signal problems in the theoretical conception of the HI scale. The contemporary literature regards hyperactivity and attention deficit as two related but separate constructs (Kuntsi et al., 2014; Willcutt et al., 2012). In this line, current diagnostic measures of attention deficit/hyperactivity disorder (ADHD), such as the one in the *Diagnostic and Statistical Manual of Mental Disorders–Fifth edition* (American Psychiatric Association, 2013) include separate conceptualizations for these traits. On one hand, attention deficit reflects the inability to focus the attention span for a sustained period of time, while on the other hand, hyperactivity–impulsivity relates to an excessive activity level combined with a lack of self-control (Garner, Marceaux, Mrug, Patterson, & Hodgens, 2010). Also, although in the past decade several studies have proposed the suitability of a general factor of ADHD (e.g., Martel, von Eye, & Nigg, 2010) a closer inspection suggests that the disorder is better represented as a multidimensional construct, rather than a single continuum (Arias, Ponce, & Núñez, 2016). Thus, the HI scale of the SDQ is bound to be problematic because it conceptualizes these two traits as being unidimensional, and by only including a few items of each (2 for hyperactivity-impulsivity and 3 for attemtion deficit), it may prevent their proper emergence as separate factors.

Regarding the robustness of the optimal five-factor ESEM structure derived and cross-validated in the current study, a subsequent Monte Carlo simulation revealed that very large samples of more than 4,000 observations would be needed to accurately recover the factor structure of the SDQ scores. In terms of the specific factors, PB was the least robust as it needed sample sizes greater than 7,000 to achieve a sufficient mean level of congruence with the structure estimated using the total sample. It is noteworthy that the optimal five-factor ESEM model included a wording method factor where the SDQ items obtained significant and nontrivial factor loadings, a result that is congruent with previous findings in the literature (e.g., Hoofs et al., 2015; McCrory & Layte, 2012; Van Roy et al., 2008). Thus, it appears that what remains of the PB factor (which contains only positive items) after extracting the wording variance from the data is not very well defined. These novel findings underscore the lack of robustness of the factorial structure underlying the SDQ self-reports.

The final step in the assessment of the factor structure of the SDQ scores involved a comparison of the ESEM results with corresponding CFA structures. In terms of model fit, there was a noticeable decrease in fit when going from the ESEM to the CFAs; indeed, many of the cross-loadings that were fixed to zero in the CFAs were significant and of nontrivial magnitude in the ESEM. However, the fit of the CFA models approximated and even surpassed conventional cutoff values established for the fit indices, so that researchers without knowledge of the ESEM results would be inclined to accept these models as providing a good-enough fit. For example, very recently Ortuño-Sierra et al. (2015) and Bøe et al. (2016) accepted five-factor CFA models of the adolescent SDQ across samples from six European countries that achieved a level of fit that was very similar to those obtained in the current study. Even more disconcerting, in the theoretical five-factor CFA with independent clusters, some of the most questionable items from the ESEM solution (the ones with the highest cross-loadings), obtained particularly high loadings in the CFA. Again, by just looking at the loadings from the CFA a researcher might mistakenly conclude that these items were strong indicators of their factors. Additionally, the factor correlations from the CFAs were considerably higher than those from the ESEM (e.g., a .37 factor correlation in the ESEM became .87 in the corresponding CFA), to the point where the discriminant validity of the factors would be questioned. This result is congruent with the literature that has shown that CFAs can grossly overestimate the factor correlations when the population model meaningfully departs from the independent clusters model (Asparouhov & Muthén, 2009; Hsu et al., 2014; Marsh et al., 2009; Schmitt & Sass, 2011). In similar fashion, Bøe et al. (2016) reported CFA factor correlations as high as .80, which at least superficially would question the suitability of CFA for their data as well (Ortuño-Sierra et al., 2015, did not report the factor correlations obtained in their study).

Regarding the internal consistency reliability of the theoretical SDQ scale scores, the results from this study showed that none achieved the minimum recommended levels of reliability (⩾.70; Cicchetti, 1994). Moreover, only the ES' scores produced a reliability estimate higher than .60. Although at first glance these results would appear to contradict recent findings of acceptable reliabilities for the adolescent SDQ (e.g., Bøe et al., 2016; Ortuño-Sierra et al., 2015), it is worth noting that these studies, along with others in the SDQ literature (e.g., Björnsdotter et al., 2013; van de Looij-Jansen et al., 2011) relied on ordinal estimators of reliability to reach these conclusions. Specifically, the ordinal alpha coefficient (Zumbo et al. 2007) used in these studies is a measure of *hypothetical* reliability, of the sum score obtained from the unobserved continuous variables that are thought to underlie the obtained discrete scores (Chalmers, 2017). As such, these reliabilities are of limited usefulness to researchers who may wish to use the SDQ scores for screening or even research purposes. Furthermore, the alpha coefficient assumes that the items do not have correlated residuals, an assumption that would be violated in the majority of studies of the SDQ. For example, whereas the observed sum scores of the HI scale achieved a reliability of .45 when taking into account its two correlated errors, the ordinal alpha coefficient produced a much higher reliability of .69.

Limitations

There are some limitations in this study that should be noted. Because this research relied on a sample from a specific region of Spain, generalizations to other cultures and languages require caution. Likewise, the current findings pertain to the adolescent self-reported SDQ, which may function differently from the parent or teacher versions, which were not evaluated here. Nevertheless, the results of this study that pertain to methodologies that have been commonly used in the SDQ literature (e.g., CFA, ordinal alpha reliability) were not too dissimilar from previous findings obtained from a diverse group of cultures and languages. Also, the very large sample size that was examined allowed for the implementation of a split-sample approach that helped ensure the stability of the findings derived from these analyses. It is also worth noting that the present study did not include external variables that could have aided the decision process that was followed to arrive at an optimal factor structure for the SDQ self-reported scores.

Practical Implications

The combined findings of the present study prevent the recommendation of the SDQ as a screening measure for the current adolescent population. First, the conceptualization of HI as a single trait is not supported by this data, previous factor analytic studies of the SDQ, or the vast literature on hyperactivity and attention deficit. Second, according to the ESEM analyses more than 25% of the SDQ items could be considered as questionable measures of their theoretical dimensions, leaving some factors with as few as three proper indicators. Third, the internal consistency reliability of the SDQ scale scores ranged from .45 to .62, which falls well below of recommended guidelines for psychological screening instruments (⩾.70; Cicchetti, 1994).

Regarding the last point, it is worth noting that the low internal consistencies of the SDQ scale scores reflect, at least partly, the inherent difficulties of trying to measure broad domains reliably with very few indicators and response options. Indeed, one of the reasons for the SDQ's popularity has been its short length, a characteristic that has been labeled as "beautiful" in relation to competing instruments with longer formats (Goodman & Scott, 1999). However, previous psychometric studies of the SDQ have consistently found poor score reliabilities. For example, in their meta-analysis of 48 studies (N = 131,223), Stone et al. (2010) found that for the SDQ Parent version four scales (all except HI) had mean internal consistency reliabilities below .70 (including two below .60). Although subsequent studies have provided higher internal consistency estimates using the ordinal alpha (Björnsdotter et al., 2013; Bøe et al., 2016; Ortuño-Sierra et al., 2015; van de Looij-Jansen et al., 2011) or ordinal omega (Gómez-Beneyto et al., 2013; Stone et al., 2013) coefficients, these measures of hypothetical reliability should be avoided, as discussed previously. In all, the findings from this study and of previous meta-analytic research indicate that trying to obtain reliably enough trait estimates with the current SDQ "small" format might be unfeasible.

The results from this study also suggest caution when interpreting the factor analytic literature of the SDQ that has relied on CFA for construct validation. As the current findings show, when an independent cluster CFA model is imposed on data that has numerous nontrivial cross-loadings, the estimated parameters are likely to be biased, potentially to a severe degree, even in cases where the CFA model has met or approximated conventional global fit criteria. This phenomenon extends beyond the SDQ, and has been documented for diverse constructs such as personality, well-being, motivation, engagement, bullying/victimization, and students' evaluations of university teaching, among others (Joshanloo, Jose, & Kielpikowski, 2017; Marsh et al., 2009; Marsh, Liem, Martin, Morin, & Nagengast, 2011; Marsh, Nagengast, & Morin, 2013; Marsh, Nagengast, Morin, Parada, et al., 2011). In light of this, we recommend that even in confirmatory applications researchers *always* estimate an ESEM model and compare its results with those obtained from the independent cluster CFA. If the fit and parameter estimates (e.g., factor correlations, main factor loadings) for the independent cluster CFA do not differ meaningfully from the corresponding ESEM, the CFA should be retained on the basis of parsimony (Marsh et al., 2014); otherwise, the ESEM model should be retained. Also, researchers can use ESEM solutions to identify large cross-loadings that could be freed in a CFA model, and then proceed to compare this modified CFA with its corresponding ESEM using the same criteria described previously. Finally, irrespective of whether an ESEM or a CFA is estimated, we encourage researchers to thoroughly inspect the local fit of their models and to at minimum report the largest SEPC for their retained models.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Supplemental Material

The online supplementary material is available at http://journals.sagepub.com/doi/suppl/10.1177/1073191118780461

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 profile*. Burlington: University of Vermont, Department of Psychiatry.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

Arias, V. B., Ponce, F. P., & Núñez, D. E. (2016). Bifactor models of attention-deficit/hyperactivity disorder (ADHD): An evaluation of three necessary but underused psychometric indexes. *Assessment*. Advance online publication. doi:10.1177/1073191116679260

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397-438. doi:10.1080/10705510903008204

Belfer, M. L. (2008). Child and adolescent mental disorders: The magnitude of the problem across the globe. *Journal of Child Psychology and Psychiatry*, *49*, 226-236. doi:10.1111/j.1469-7610.2007.01855.x

Björnsdotter, A., Enebrink, P., & Ghaderi, A. (2013). Psychometric properties of online administered parental Strengths and Difficulties Questionnaire (SDQ), and normative data based on combined online and paper-and-pencil administration. *Child and Adolescent Psychiatry and Mental Health*, *7*, 40. doi:10.1186/1753-2000-7-40

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*, e0152202. doi:10.1371/journal.pone.0152202

Caci, H., Morin, A. J., & Tran, A. (2015). Investigation of a bifactor model of the Strengths and Difficulties Questionnaire. *European Child & Adolescent Psychiatry*, *24*, 1291-1301. doi:10.1007/s00787-015-0679-3

Capron, C., Thérond, C., & Duyme, M. (2007). Psychometric properties of the French version of the self-report and teacher Strengths and Difficulties Questionnaire (SDQ). *European Journal of Psychological Assessment*, *23*, 79-88. doi:10.1027/1015-5759.23.2.79

Chalmers, R. P. (2017). On misconceptions and the limited usefulness of ordinal alpha. *Educational and Psychological Measurement*. Advance online publication. doi:10.1177/0013164417727036

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*, 464-504. doi:10.1080/10705510701301834

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*, 63-74. doi:10.1177/1073191114568301

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*, 284-290. doi:10.1037/1040-3590.6.4.284

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155-159. doi:10.1037/0033-2909.112.1.155

Cook, A., Spinazzola, J., Ford, J., Lanktree, C., Blaustein, M., Cloitre, M., . . . Mallah, K. (2017). Complex trauma in children and adolescents. *Psychiatric Annals*, *35*, 390-398. doi:10.3928/00485713-20050501-05

Du, Y., Kou, J., & Coghill, D. (2008). The validity, reliability and normative scores of the parent, teacher and self report versions of the Strengths and Difficulties Questionnaire in China. *Child and Adolescent Psychiatry and Mental Health*, *2*, 8. doi:10.1186/1753-2000-2-8

Elander, J., & Rutter, M. (1996). Use and development of the Rutter parents' and teachers' scales. *International Journal of Methods in Psychiatric Research*, *6*, 63-78. doi:10.1002/(SICI)1234-988X(199607)6:2<63::AID-MPR151>3.3.CO;2-M

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*, 195-212. doi:10.3758/s13428-017-0862-1

Epskamp, S., Maris, G., Waldorp, L. J., & Borsboom, D. (in press). Network psychometrics. In P. Irwing, D. Hughes, & T. Booth (Eds.), *Handbook of psychometrics*. New York, NY: Wiley.

Essau, C. A., Olaya, B., Anastassiou-Hadjicharalambous, X., Pauli, G., Gilvarry, C., Bray, D., . . . Ollendick, T. H. (2012). Psychometric properties of the Strength and Difficulties Questionnaire from five European countries. *International Journal of Methods in Psychiatric Research*, *21*, 232-245. doi:10.1002/mpr.1364

Fox-Wasylyshyn, S. M., & El-Masri, M. M. (2005). Handling missing data in self-report measures. *Research in Nursing & Health*, *28*, 488-495. doi:10.1002/nur.20100

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). Retrieved from http://pareonline.net/getvn.asp?v=17&n=3

García, P., Goodman, R., Mazaira, J., Torres, A., Rodríguez-Sacristán, J., Hervas, A., & Fuentes, J. (2000). El cuestionario de Capacidades y Dificultades [The Strengths and Difficulties Questionnaire]. *Revista de Psiquiatría Infanto-Juvenil*, *1*, 12-17.

Garner, A. A., Marceaux, J. C., Mrug, S., Patterson, C., & Hodgens, B. (2010). Dimensions and correlates of attention deficit/hyperactivity disorder and sluggish cognitive tempo. *Journal of Abnormal Child Psychology*, *38*, 1097-1107. doi:10.1007/s10802-010-9436-8

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods*, *18*, 454-474. doi:10.1037/a0030005

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via Monte Carlo simulation. *Psychological Methods*, *21*, 93-111. doi:10.1037/met0000064

Golino, H. F. (2017). *EGA package*. Retrieved from github.com/hfgolino/EGA

Golino, H. F., & Epskamp, S. (2017). Exploratory graph analysis: A new approach for estimating the number of dimensions in psychological research. *PLoS ONE*, *12*, e0174035. doi:10.1371/journal.pone.0174035

Gómez-Beneyto, M., Nolasco, A., Moncho, J., Pereyra-Zamora, P., Tamayo-Fonseca, N., Munarriz, M., . . . Girón, M. (2013). Psychometric behaviour of the Strengths and Difficulties Questionnaire (SDQ) in the Spanish national health survey 2006. *BMC Psychiatry*, *13*, 1-10. doi:10.1186/1471-244X-13-95

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581-586. doi:10.1111/j.1469-7610.1997.tb01545.x

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*, 1337-1345. doi:10.1097/00004583-200111000-00015

Goodman, R., & Scott, S. (1999). Comparing the Strengths and Difficulties Questionnaire and the Child Behavior Checklist: Is small beautiful? *Journal of Abnormal Child Psychology*, *27*, 17-24. doi:10.1023/A:1022658222914

Gore, F. M., Bloem, P. J., Patton, G. C., Ferguson, J., Joseph, V., Coffey, C., . . . Mathers, C. D. (2011). Global burden of disease in young people aged 10–24 years: A systematic analysis. *Lancet*, *377*, 2093-2102. doi:10.1016/S0140-6736(11)60512-6

Guay, F., Morin, A. J., Litalien, D., Valois, P., & Vallerand, R. J. (2014). Application of exploratory structural equation modeling to evaluate the Academic Motivation Scale. *Journal of Experimental Education*, *83*, 51-82. doi:10.1080/00220973.2013.876231

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Upper Saddle River, NJ: Pearson.

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods*, *7*, 191-205. doi:10.1177/1094428104263675

He, J. P., Burstein, M., Schmitz, A., & Merikangas, K. R. (2013). The Strengths and Difficulties Questionnaire (SDQ): The factor structure and scale validation in US adolescents. *Journal of Abnormal Child Psychology*, *41*, 583-595. doi:10.1007/s10802-012-9696-6

Hoofs, H., Jansen, N. W. H., Mohren, D. C. L., Jansen, M. W. J., & Kant, I. J. (2015). The context dependency of the self-report version of the Strength and Difficulties Questionnaire (SDQ): A cross-sectional study between two administration settings. *PLoS ONE*, *10*, e0120930. doi:10.1371/journal.pone.0120930

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179-185. doi:10.1007/BF02289447

Hsu, H. Y., Skidmore, S. T., Li, Y., & Thompson, B. (2014). Forced zero cross-loading misspecifications in measurement component of structural equation models: Beware of even "small" misspecifications. *Methodology*, *10*, 138-152. doi:10.1027/1614-2241/a000084

Hu, L. T., & 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

Joshanloo, M., Jose, P. E., & Kielpikowski, M. (2017). The value of exploratory structural equation modeling in identifying factor overlap in the Mental Health Continuum-Short Form (MHC-SF): A study with a New Zealand sample. *Journal of Happiness Studies*, *18*, 1061-1074. doi:10.1007/s10902-016-9767-4

Kieling, C., Baker-Henningham, H., Belfer, M., Conti, G., Ertem, I., Omigbodun, O., . . . Rahman, A. (2011). Child and adolescent mental health worldwide: Evidence for action. *Lancet*, *378*, 1515-1525. doi:10.1016/S0140- 6736(11)60827-1

Kim, M. H., Ahn, J. S., & Min, S. (2015). Psychometric properties of the self-report version of the Strengths and Difficulties Questionnaire in Korea. *Psychiatry Investigation*, *12*, 491-499. doi:10.4306/pi.2015.12.4.491

Koskelainen, M., Sourander, A., & Vauras, M. (2001). Self-reported strengths and difficulties in a community sample of Finnish adolescents. *European Child & Adolescent Psychiatry*, *10*, 180-185. doi:10.1007/s007870170024

Kuntsi, J., Pinto, R., Price, T. S., van der Meere, J. J., Frazier-Wood, A. C., & Asherson, P. (2014). The separation of ADHD inattention and hyperactivity-impulsivity symptoms: Pathways from genetic effects to cognitive impairments and symptoms. *Journal of Abnormal Child Psychology*, *42*, 127-136. doi:10.1007/s10802-013-9771-7

Liu, S. K., Chien, Y. L., Shang, C. Y., Lin, C. H., Liu, Y. C., & Gau, S. S. F. (2013). Psychometric properties of the Chinese version of Strength and Difficulties Questionnaire. *Comprehensive Psychiatry*, *54*, 720-730. doi:10.1016/j.comppsych.2013.01.002

Lorenzo-Seva, U., & ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57-64. doi:10.1027/1614-2241.2.2.57

Lundh, L., Wångby-Lundh, M., & Bjärehed, J. (2008). Self-reported emotional and behavioral problems in Swedish 14 to 15-year-old adolescents: A study with the self-report version of the Strengths and Difficulties Questionnaire. *Scandinavian Journal of Psychology*, *49*, 523-532. doi:10.1111/j.1467-9450.2008.00668.x

Mansbach-Kleinfeld, I., Apter, A., Farbstein, I., Levine, S. Z., & Ponizovsky, A. M. (2010). A population-based psychometric validation study of the Strengths and Difficulties Questionnaire–Hebrew version. *Child and Neurodevelopmental Psychiatry*, *1*, 1-12. doi:10.3389/fpsyt.2010.00151

Marsh, H. W., Hau, K. T., & 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. doi:10.1207/s15328007sem1103_2

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J., & 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*, 322-346. doi:10.1177/0734282911406657

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85-110. doi:10.1146/annurev-clinpsy-032813-153700

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling*, *16*, 439-476. doi:10.1080/10705510903008220

Marsh, H. W., Nagengast, B., & Morin, A. J. (2013). 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*, 1194-1218. doi:10.1037/a0026913

Marsh, H. W., Nagengast, B., Morin, A. J., Parada, R. H., Craven, R. G., & Hamilton, L. R. (2011). Construct validity of the multidimensional structure of bullying and victimization: An application of exploratory structural equation modeling. *Journal of Educational Psychology*, *103*, 701-732. doi:10.1037/a0024122

Marsh, H. W., Scalas, L. F., & Nagengast, B. (2010). Longitudinal tests of competing factor structures for the Rosenberg Self-Esteem Scale: Traits, ephemeral artifacts, and stable response styles. *Psychological Assessment*, *22*, 366-381. doi:10.1037/a0019225

Martel, M. M., von Eye, A., & Nigg, J. T. (2010). Revisiting the latent structure of ADHD: Is there a "g" factor? *Journal of Child Psychology and Psychiatry*, *51*, 905-914. doi:10.1111/j.1469-7610.2010.02232.x

Mathai, J., Anderson, P., & Bourne, A. (2004). Comparing psychiatric diagnoses generated by the Strengths and Difficulties Questionnaire with diagnoses made by clinicians. *Australian & New Zealand Journal of Psychiatry*, *38*, 639-643.

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*, 344-362. doi:10.1037/1082-989X.11.4.344

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

Mellor, D. (2005). Normative data for the Strengths and Difficulties Questionnaire in Australia. *Australian Psychologist*, *40*, 215-222. doi:10.1080/00050060500243475

Mellor, D., & Stokes, M. (2007). The factor structure of the Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment*, *23*, 105-112. doi:10.1027/1015-5759.23.2.105

Merikangas, K. R., He, J. P., Burstein, M., Swanson, S. A., Avenevoli, S., Cui, L., . . . Swendsen, J. (2010). Lifetime prevalence of mental disorders in US adolescents: Results from the National Comorbidity Survey Replication–Adolescent Supplement (NCS-A). *Journal of the American Academy of Child & Adolescent Psychiatry*, *49*, 980-989. doi:10.1016/j.jaac.2010.05.017

Merikangas, K. R., Nakamura, E. F., & Kessler, R. C. (2009). Epidemiology of mental disorders in children and adolescents. *Dialogues in Clinical Neuroscience*, *3*, 7-20.

Morin, A. J., Arens, A. K., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*, 116-139. doi:10.1080/10705511.2014.961800

Niclasen, J., Skovgaard, A. M., Andersen, A. N., Sømhovd, M. J., & Obel, C. (2013). A confirmatory approach to examining the factor structure of the Strengths and Difficulties Questionnaire (SDQ): A large scale cohort study. *Journal of Abnormal Child Psychology*, *41*, 355-365. doi:10.1007/s10802-012-9683-y

Ortuño-Sierra, J., Fonseca-Pedrero, E., Aritio-Solana, R., Velasco, A. M., de Luis, E. C., Schumann, G., . . . Bokde, A. (2015). New evidence of factor structure and measurement invariance of the SDQ across five European nations. *European Child & Adolescent Psychiatry*, *24*, 1523-1534. doi:10.1007/s00787-015-0729-x

Palmieri, P. A., & Smith, G. C. (2007). Examining the structural validity of the Strengths and Difficulties Questionnaire (SDQ) in a U.S. sample of custodial grandmothers. *Psychological Assessment*, *19*, 189-198. doi:10.1037/1040-3590.19.2.189

Patalay, P., Hayes, D., Deighton, J., & Wolpert, M. (2016). A comparison of paper and computer administered Strengths and Difficulties Questionnaire. *Journal of Psychopathology and Behavioral Assessment*, *38*, 242-250. doi:10.1007/s10862-015-9507-9

Patel, V., Flisher, A. J., Hetrick, S., & McGorry, P. (2007). Mental health of young people: A global public-health challenge. *Lancet*, *369*, 1302-1313. doi:10.1016/S0140-6736(07)60368-7

Percy, A., McCrystal, P., & Higgins, K. (2008). Confirmatory factor analysis of the adolescent self-report Strengths and Difficulties Questionnaire. *European Journal of Psychological Assessment*, *24*, 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*, 345-365. doi:10.1111/jcpp.12381

Pons, P., & Latapy, M. (2006). Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, *10*, 191-218.

Rescorla, L., Achenbach, T. M., Ivanova, M. Y., Dumenci, L., Almqvist, F., Bilenberg, N., . . . Erol, N. (2007). Epidemiological comparisons of problems and positive qualities reported by adolescents in 24 countries. *Journal of Consulting and Clinical Psychology*, *75*, 351-358. doi:10.1037/0022-006X.75.2.351

Revelle, W. (2017). *psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University.

Reynolds, C. R., & Kamphaus, R. W. (1992). *Behavior assessment system for children*. Circle Pines, MN: American Guidance Service.

Rhemtulla, M., Brosseau-Liard, P. É., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354-373. doi:10.1037/a0029315

Ruchkin, V., Jones, S., Vermeiren, R., & Schwab-Stone, M. (2008). The Strengths and Difficulties Questionnaire: The self-report version in American urban and suburban youth. *Psychological Assessment*, *20*, 175-182. doi:10.1037/1040-3590.20.2.175

Saris, W. E., Satorra, A., & van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications? *Structural Equation Modeling*, *16*, 561-582. doi:10.1080/10705510903203433

Schmitt, T. A. (2011). Current methodological considerations in exploratory and confirmatory factor analysis. *Journal of Psychoeducational Assessment*, *29*, 304-321. doi:10.1177/0734282911406653

Schmitt, T. A., & Sass, D. A. (2011). Rotation criteria and hypothesis testing for exploratory factor analysis: Implications for factor pattern loadings and interfactor correlations. *Educational and Psychological Measurement*, *71*, 95-113. doi:10.1177/0013164410387348

Smedje, H., Broman, J., Hetta, J., & von Knorring, A. (1999). Psychometric properties of a Swedish version of the "Strengths and Difficulties Questionnaire." *European Child & Adolescent Psychiatry*, *8*, 63-70. doi:10.1007/s007870050086

Stevanovic, D., Urbán, R., Atilola, O., Vostanis, P., Balhara, Y. S., Avicenna, M., ... & Petrov, P. (2015). Does the Strengths and Difficulties Questionnaire–self report yield invariant measurements across different nations? Data from the International Child Mental Health Study Group. *Epidemiology and Psychiatric Sciences*, *24*(4), 323-334. doi:10.1017/S2045796014000201

Stone, L., Otten, R., Engels, R., Vermulst, A., & Janssens, J. (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*, 254-274. doi:10.1007/s10567-010-0071-2

Stone, L. L., Otten, R., Ringlever, L., Hiemstra, M., Engels, R. C., Vermulst, A. A., & Janssens, J. M. (2013). The parent version of the Strengths and Difficulties Questionnaire: Omega as an alternative to alpha and a test for measurement invariance.

European Journal of Psychological Assessment, *29*, 44-50. doi:10.1027/1015-5759/a000119

Tobia, V., Gabriele, M. A., & Marzocchi, G. M. (2013). The Italian version of the Strengths and Difficulties Questionnaire (SDQ)—Teacher: Psychometric properties. *Journal of Psychoeducational Assessment*, *31*, 493-505. doi:10.1177/0734282912473456

Tucker, L. R. (1951). *A method for synthesis of factor analysis studies* (Personnel Research Section Report No. 984). Washington, DC: Department of the Army.

van de Looij-Jansen, P. M., Goedhart, A. W., de Wilde, E. J., & Treffers, P. D. (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*, 127-144. doi:10.1348/014466510X498174

Van Roy, B., Grøholt, B., Heyerdahl, S., & Clench-Aas, J. (2006). Self-reported strengths and difficulties in a large Norwegian population 10–19 years: Age and gender specific results of the extended SDQ-questionnaire. *European Child & Adolescent Psychiatry*, *15*, 189-198. doi:10.1007/s00787-005-0521-4

Van Roy, B., Veenstra, M., & Clench-Aas, J. (2008). Construct validity of the five-factor Strengths and Difficulties Questionnaire (SDQ) in pre-, early, and late adolescence. *Journal of Child Psychology and Psychiatry*, *49*, 1304-1312. doi:10.1111/j.1469-7610.2008.01942.x

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología*, *33*, 755-782.

Warnick, E. M., Bracken, M. B., & Kasl, S. (2008). Screening efficiency of the Child Behavior Checklist and Strengths and Difficulties Questionnaire: A systematic review. *Child and Adolescent Mental Health*, *13*, 140-147. doi:10.1111/j.1475-3588.2007.00461.x

Weijters, B., Baumgartner, H., & Schillewaert, N. (2013). Reversed item bias: An integrative model. *Psychological Methods*, *18*, 320-334. doi:10.1037/a0032121

Whittaker, T. A. (2012). Using the modification index and standardized expected parameter change for model modification. *Journal of Experimental Education*, *80*, 26-44. doi:10.1080/00220973.2010.531299

Widaman, K. F. (2006). III. Missing data: What to do with or without them. *Monographs of the Society for Research in Child Development*, *71*, 42-64.

Willcutt, E. G., Nigg, J. T., Pennington, B. F., Solanto, M. V., Rohde, L. A., Tannock, R., . . . Lahey, B. B. (2012). Validity of DSM-IV attention deficit/hyperactivity disorder symptom dimensions and subtypes. *Journal of Abnormal Psychology*, *121*, 991-1010.

Woods, C. M. (2006). Careless responding to reverse-worded items: Implications for confirmatory factor analysis. *Journal of Psychopathology and Behavioral Assessment*, *28*, 186. doi:10.1007/s10862-005-9004-7

Yang, Y., & Green, S. B. (2015). Evaluation of structural equation modeling estimates of reliability for scales with ordered categorical items. *Methodology*, *11*, 23-34. doi:10.1027/1614-2241/a000087

Yao, S., Zhang, C., Zhu, X., Jing, X., McWhinnie, C. M., & Abela, J. R. (2009). Measuring adolescent psychopathology: Psychometric properties of the self-report Strengths and Difficulties Questionnaire in a sample of Chinese adolescents. *Journal of Adolescent Health*, *45*, 55-62. doi:10.1016/j.jadohealth.2008.11.006

Zumbo, B. D., Gadermann, A. M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for Likert rating scales. *Journal of Modern Applied Statistical Methods*, *6*, 21-29. doi:10.22237/jmasm/1177992180