Is it time to revise the SDQ? The Psychometric Evaluation of the Strengths and Difficulties Questionnaire

Reeta Kankaanpää[1], Pertti Töttö[2], Raija-Leena Punamäki[1] & Kirsi Peltonen[3]

[1] Faculty of Social Sciences / Psychology, Tampere University

[2] Faculty of Social Sciences and Business Studies, University of Eastern Finland

[3] INVEST Research Flagship Center, University of Turku, Finland

Author Note

Reeta Kankaanpää https://orcid.org/0000-0001-9111-7076

Reeta Kankaanpää is now affiliated both in [1] Tampere University and [3] INVEST Research Flagship Center, University of Turku, Finland.

This study is part of the RefugeesWellSchool (RWS) study Horizon 2020 research project (No 754849; ISRCTN64245549). Authors have no conflicts of interest to disclosure.

Correspondence concerning this article should be addressed to Reeta Kankaanpää, Faculty of Social Sciences / Psychology, FIN-33014 Tampere University, Finland. E-Mail: reeta.kankaanpaa@tuni.fi.

All data have been made publicly available at the Finnish Social Science Data Archive AILA, and links to data, and all analysis code for this article can be accessed at https://osf.io/vnw84/. This study was not preregistered.

Author contribution statement with CRediT guidelines

Reeta Kankaanpää, MSSc (Conceptualization: Lead; Formal analysis: Lead; Methodology: Lead; Visualization: Lead; Writing – original draft: Lead; Writing – review & editing: Lead) Pertti Töttö, PhD (Conceptualization: Supporting; Writing – review & editing: Supporting) Raija-Leena Punamäki, PhD (Conceptualization: Supporting; Funding acquisition: Lead; Resources: Lead; Supervision: Equal; Writing –

review & editing: Supporting) Kirsi Peltonen, PhD (Conceptualization: Supporting; Supervision: Equal; Writing – review & editing: Supporting)

© 2023, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/pas0001265

Abstract

Despite the wide use of the Strengths and Difficulties Questionnaire (SDQ) to assess adolescent mental health, its psychometric functionality is still under debate. This study investigated the structural validity and reliability of the SDQ scores, and the resemblance of the SDQ sum scores and factor scores. Factor one-dimensionality and competing multi-factor structures were tested against data. With the best acceptable models, measurement invariance was tested between genders and over time. Subscale reliability and correspondence between subscale sum scores and factor scores were estimated. The nationally representative self-report data from 23,980 Finnish early (12–13 years) and mid- (15–16 years) adolescents (50.4 % girls) was collected from two cohorts in 2008 and 2013. The results showed that among early adolescents, the revised SDQ with a controlled method effect had an excellent fit. In contrast, none of the tested models had an acceptable fit among the mid-adolescents. Among early adolescents, strong measurement invariance was achieved between genders and over time. Three of the five subscales were one-dimensional, and all subscales had low reliability. The resemblance between the subscale sum scores and factor scores was alarmingly low. Researchers should be cautious when using the SDQ Total Difficulties sum score or the subscale scores as they may be substantially biased, and practitioners should desist from using the SDQ as a screening tool in its current form. This study strongly supports the revision of the SDQ. In line with previous findings, we suggest rewording the worst functioning items and revising the reverse-worded difficulties items.

Keywords: the psychometric properties, the Strengths and Difficulties Questionnaire (SDQ), adolescents, structural validity, reliability, sum score and factor score resemblance

Public significance statement:

>> The self-reported SDQ contains method effects which can and should be controlled when the SDQ is used in research, and more research is needed to guarantee the reliable use of the SDQ sum scores for assessing adolescent mental health, because the sum scores in their current form may be substantially biased.

Is It Time to Revise the SDQ? The Psychometric Evaluation of the Strengths and Difficulties Questionnaire

Adolescents are at risk for mental health problems due to rapid psychological, social, and biophysiological changes. Research shows that approximately every sixth adolescent exhibits emotional or behavioral problems (Barkmann & Schulte-Markwort, 2012; Philipp et al., 2018). If not adequately treated, adolescents' mental health problems may accumulate in adulthood (Clayborne et al., 2019; Merikukka et al., 2018). Appropriate and timely prevention and effective treatments require adequate instruments to screen and assess the emotional and behavioral problems. One of the most well-known instruments is the Strengths and Difficulties Questionnaire (SDQ) (R. Goodman, 2001). However, the results about the SDQ's psychometric properties have been conflicting and it is debatable whether the SDQ, in its current form, should be used to assess adolescent mental health (Duinhof et al., 2019; Garrido et al., 2020; Vugteveen et al., 2020). This study offers a comprehensive psychometric analysis of the SDQ by examining the validity of the factor structure, estimating the reliability of the subscales using alpha, ordinal alpha, and omega coefficients, and estimating the sum score and factor score resemblance.

The SDQ Assessing Adolescent Mental Health

The SDQ is a brief screening questionnaire for assessing the mental health of children and adolescents. It consists of 15 negative and 10 positive items that are meant to address five distinct dimensions, each with five items: "emotional problems" (EP), "conduct problems" (CP), "hyperactivity-inattention" (HA), "peer problems" (PP), and "prosocial behavior" (PB). Five of the positive items form the prosocial behavior scale and the other five items are dispersed on the four Difficulties scales measuring the absence of the problems. The conduct problems scale contains one positive item ("obedient"), the peer problems scale contains two positive items ("friend" and "popular"), and the hyperactivity-inattention scale contains two positive items ("reflective" and "persistent"). Each dimension has five items. The item wording, labels, and names can be found in the Appendix, Table S1.

The SDQ is commonly used in clinical settings and community studies for screening and assessing the mental health of adolescents. As a screening instrument, the SDQ is used to select adolescents for further evaluation, thus providing information for diagnosing psychopathology or disorders. The emotional problems scale is thought to indicate two overlapping disorders: depression and anxiety. The conduct problems scale indicates oppositional defiant disorder (ODD) and the hyperactivity-inattention scale indicates attention-deficit hyperactivity disorder (ADHD). The prosocial behavior scale is based on a latent trait called prosocial behavior (Davidov et al., 2016; Weir & Duveen, 1981). The peer problems scale does not indicate any specific disorder.

In population screenings, the four subscales measuring difficulties are recommended to be treated as pairs, that are then called “Internalizing problems” (EP+PP), and “Externalizing problems” (CP+HA) (A. Goodman et al., 2010). In research, the SDQ is mostly used as the Total Difficulties sum score, where the four difficulties subscales are summed together. The SDQ can be collected as a self-report, or as a report by a caregiver or teacher. Here, we focus on the self-report version of the SDQ.

Extensive literature is available on the structural validity and reliability of the self-reported SDQ scores. Table 1 presents the earlier studies with the information about the country they are conducted, the age of participants, and the sample sizes. Table 2 reports the psychometric results of the studies about structural validity, reliability, and the use of the Total Difficulties Score.

[Table 1.]

[Table 2.]

Structural Validity of the SDQ Scores

Several previous studies support the original five-factor structure. Other studies suggest structural validity to be based on three, four, or six factors instead of the original five. Half of the studies reported findings against the five-factor structure showing three main statistical challenges in the structural validity of the SDQ scores: cross-loadings, residual covariances, and low loadings. Removing the reverse-worded items, using a method factor, and adding residual covariances has been suggested to solve the abovementioned challenges.

First, reverse-worded difficulties items tend to load on the prosocial behavior factor containing positively worded items (Hoofs et al., 2015; Vugteveen et al., 2020) and some studies have solved the cross-loading by removing the reverse-worded items from the model (Duinhof et al., 2019; Essau et al., 2012). It may nevertheless be problematic for reliability because the SDQ is already a short instrument, and the removal of items makes it even more unreliable. Table 2 shows that six studies have tried to solve cross-loadings by fixing the method-related variance using different forms of method factor. Most of these studies allow the method factor to correlate with the theoretic factors despite that the two most known method factor models do not allow correlations between the theoretic factors and the method factor. The two models are the method factor model by Campbell and Fiske (1959) and updated by Eid (2000), and the method factor model by Maydeu-Olivares and Coffman (2006). In both versions, the method factor is expected to be orthogonal to theoretic factors. The difference is in which items are set to load on the method factor. In Campbell and Fiske's (1959) and Eid's (2000) version, only those items which are thought to be affected by the method, and in Maydeu-Olivares and Coffman's (2006) version, all items are set to load on the method factor. Additionally, factor loadings on the method factor are constrained equally. Deciding about the best method factor model is still under debate (Nieto et al., 2021). No previous study on the structural validity of the SDQ scores has set the method factor orthogonal to theoretic factors when using the method factor model by Campbell and Fiske and Eid and only one study used the RIIFA to account for a method effect in the SDQ (Garrido et al., 2020) based on Maydeu-Olivares and Coffman.

Second, many studies have reported improving the measurement model by adding residual covariances. Residual covariances are statistically equivalent to additional factors orthogonal to original factors, and thus they require meaningful theoretical interpretation in the same way as the original factors. Recent studies have started to talk about redundant items and suggested rewording or removing them (Garrido et al., 2020; Ribeiro Santiago et al., 2021; Santiago et al., 2021). Redundancy of the items is however meaningful only if the SDQ is modelled as a symptom network. In the factor analytic approach, the ideal model consists of nearly identical items that are included in the measurement model only to increase

reliability. Thus, in the factor analytic approach, residual covariances are more a sign of multidimensionality that should be carefully investigated.

Third, low loadings have appeared in several studies. Some studies have recommended removing the item “obedient” for not adequately reflecting pathological behavior as some amount of disobedience is an essential part of normative development (Bøe et al., 2016). The problems of low loadings in “adults” can result from the item being outdated and in “persistent” from being very close to other items in the hyperactivity-inattention scale, such as “distractible” or “reflective”. On the other hand, “restless” and “fidgety” are very close to each other too. In fact, the hyperactivity-inattention factor may consist of two subfactors.

Cross-loadings, residual covariances, and low loadings indicate multidimensionality within the factor structure and complicate interpretations of the structural validity of the SDQ. Our study investigates the structural validity by testing the one-dimensionality of subscales, testing two kinds of method factors to the multi-factor structure, and removing the lowest loading items.

Measurement invariance of the SDQ Scores

A sign of validity and high quality of a questionnaire is that it functions similarly across different groups, such as among boys and girls. Statistically that is indicated by the measurement invariance. As shown in Table 2, several studies have reported results on the measurement invariance of the SDQ. A few studies have achieved measurement invariance between gender but only one of the studies used the original measurement model (Yao et al., 2009). Measurement invariance across age groups has been shown in some studies. However, all these studies added modifications in the measurement model before invariance testing. Measurement invariance should be tested only after finding an adequate measurement model in one group. For instance, Koskelainen (2001) reported clear differences in the measurement model between genders without invariance testing. Considering that only one of the previous studies achieved measurement invariance using the original structure, there is thus a need for the statistically sophisticated testing of measurement invariance for both gender and age, which is the contribution of the current study.

Reliability of the SDQ Scores

The estimation of reliability is an essential part of the psychometric evaluation of the SDQ, and a low reliability in scales should be considered as a warning sign. Table 2 presents that a substantial number of studies have reported low reliability for the SDQ subscales. Low alpha estimates in the SDQ subscales may result from ordinal items, nonnormal distributions, or multidimensional scales. Studies using reliability coefficients such as ordinal alpha and omega that can account for the ordinality of the items have reported higher reliability estimates compared to studies using Cronbach's alpha. In the most studies on SDQ scales, the reliability has been based only on a few statistical estimates, commonly alpha. It would be informative to provide reliability estimates using a statistical approach that can account for the ordinality of the items and the method-related variance. Our study provides reliability estimates from several different statistical approaches, including ordinal alpha and omega.

Many studies have reported that the reliability of the SDQ subscales is low but sufficient for the Total Difficulties sum score. However, it is then worth asking what the Total Difficulties sum score measures: Does it measure some real difficulties or systematic error variance? If the Total Difficulties sum score contains a substantial amount of method-related error, the risk for misclassification is greatly increased. Ribeiro-Santiago (Santiago et al., 2021) provided a thought-provoking example of how a reliability coefficient of .65 would imply that around 40% of all true positives will be misclassified (Charter & Feldt, 2001). Hence, the use of sum scores with a low reliability has been strongly discouraged in clinical screening in which important decisions will be made on adolescents' lives (Charter, 2003).

In the Total Difficulties sum score, the 20 Difficulties items are simply summed together. According to Table 2, several studies have supported the use of the Total Difficulties sum score and subscale sum scores. Some studies have supported the use of the Total Difficulties sum score but not the use of the subscale sum scores. Further, some studies have recommended not using any SDQ sum scores as such or at least modifying the scales before use. Despite the methodological shortcomings in the SDQ scores' structural validity reviewed above, the SDQ is used as a Total Difficulties sum score in screening and as an outcome in observational and experimental studies (Lesinskiene et al., 2018; Peltonen et al., 2022).

As an alternative to sum scores, in research settings, the SDQ scales could be used as factor scores, or as a latent variable model. In the factor scores, each item contributes to the score with a unique weight, whereas in the latent variable model, items have unique weights, and the measurement error is controlled so that the latent variable has full reliability. Sum scores can differ greatly from the tested measurement model if the model is modified, or factor loadings vary. Only one study has shown how strongly the SDQ sum scores and factor scores correlate (Vugteveen et al., 2020). Correlations varied between .900 and .976, indicating 81–95 percent common variation. In other words, 5 to 19 percent of the variance differed between the sum scores and factor scores showing reasonable resemblance. Following the findings by Vugteveen (Vugteveen et al., 2020), we estimate the resemblance between subscale sum scores and factor scores.

Aims of the Current Study

This study aims to offer a comprehensive psychometric analysis of the SDQ using a representative sample of Finnish adolescents. The research tasks are as follows: to examine structural validity, measurement invariance, and reliability of the SDQ scores, and to estimate how well the subscale sum scores and factor scores resemble each other.

First, we evaluate the structural validity of the SDQ by testing subscale one-dimensionality and essential τ-equivalence of the factor loadings in the subscales of emotional problems, conduct problems, hyperactivity-inattention, peer problems, and prosocial behavior. We test the structural validity of four competing factor structures: the original five-factor structure, the four-factor structure known as the Total Difficulties (sum score) structure, a revised SDQ with a method factor model, and a revised SDQ with a RIIFA factor model. With the best acceptable factor structure, we test measurement invariance across genders, age groups, and over time. Second, we estimate the reliability of the SDQ subscales by using various estimates: alpha, ordinal alpha, and omega coefficients. Third, we estimate the resemblance between the subscale sum scores and factor scores.

Methods

Data

We analyze two data sets collected as part of the Child Victim Surveys in 2008 and 2013 at Finnish schools using multistage probability sampling by Statistics Finland (Ellonen et al., 2008, 2013). The data sets contain multiple measures concerning children's and adolescents' experiences and wellbeing. The data is stored and available for research in the Finnish Social Science Data Archive. This study utilizes data only for the SDQ. The sampling unit in the surveys was school class. The samples were stratified according to province, municipality type and school size. Schools requested students from one to three classes to participate in the study depending on the school size. The data consist of responses from students who were early adolescents (12–13 years old) or mid-adolescents (15–16 years old). In 2008, 88 percent of the early adolescent students and 64 percent of the mid adolescent students responded. The total number of respondents was 13,459. There was no non-response bias (Ellonen et al., 2008). In 2013, there were 11,364 respondents and the response bias was not reported. Data collection was carried out during a class held in a computer lab with a teacher present. The questionnaire was published on the research project website. The website included instructions for respondents and some additional information. The data sets were divided into eight distinct groups. The variables used for creating groups were gender (male, female), age (early adolescents, mid-adolescents), and year of data collection. Below, we refer to the groups based on gender, stage of adolescence, and time of collection: time 1 (2008) and time 2 (2013).

The demographic characteristics of the groups are reported in the Appendix Table S2. The chi-squared test showed a statistically significant difference in parental education, subjective income, proportion of reported mental health problems, and learning difficulty between the two times. The test is however sensitive to the large sample size. Therefore, we provided Cramer's V estimates to indicate the effect size. Parental education, a self-reported mental health problem, and a learning difficulty had a slightly greater than a small effect. This indicates that the groups had a small to medium size difference in their mental health, learning difficulty, and parents' education. The mid-adolescents in particular reported

having a mental health problem. Having a learning difficulty and higher parental education was more prominent at time 2 than at time 1.

This study was not preregistered. Since the study is conducted using secondary data which has already been approved by the ethics committee, the study did not involve ethics committee review or approval. All data have been made publicly available at the Finnish Social Science Data Archive AILA and can be accessed via the links mentioned in the references. All analysis codes can be accessed at OSF: https://osf.io/vnw84/?view_only=cdfa4c8298ca4949a1e67c3e882486b6.

Measures

Strengths and Difficulties Questionnaire (SDQ)

The SDQ is a 25-item brief screening tool that measures adolescents' behaviors, emotions, and relationships (R. Goodman, 1997, 2001). In this study, we evaluated the SDQ self-report version for adolescents aged 12–16 years. Each item is rated on a three-point scale: *not true*, *somewhat true*, or *certainly true*. Fifteen Difficulties items are worded negatively, and five Difficulties items and five prosocial behavior items are worded positively. The subscales are called emotional problems, conduct problems, hyperactivity-inattention, peer problems, and prosocial behavior. Each of the five subscales contains five items.

Analytic Strategy

The statistical analyses were conducted with R software (R Core Team, 2021), R packages *lavaan* (Rosseel, 2012), and *semTools* (Jorgensen et al., 2022). CFA models were estimated using a DWLS estimator suitable for ordered variables. The lavaan estimation of categorical variables does not support full information maximum likelihood, therefore only complete data were used in the analysis. In the OSF webpage, the missing value tables are reported by groups. The number of missing values varied between 3 and 114 per item. The overall average of missing values was 37 per item.

First, we tested the one-dimensionality of each scale. If a scale was found to be one-dimensional, we tested whether the factor loadings were essentially τ-equivalent. Essential τ-equivalence was tested by setting factor loadings as equal using the same label for all.

Second, we tested the four competing multi-factor structures of the SDQ: the original five-factor structure, the Total Difficulties four-factor structure, the revised four-factor structure with a method factor, and the revised four-factor structure with a RIIFA factor. The four-factor structures contained the Difficulties factors (emotional problems, conduct problems, hyperactivity-inattention, peer problems). After inspecting residuals, we tested a model where an item with the least explained variance was removed. This resulted in removing two items ("obedient" from the conduct problems scale and "adults" from the peer problems scale), yet all scales were left with four or more items. We added an orthogonal method factor according to Campbell and Fiske (Campbell & Fiske, 1959) and Eid (Eid, 2000) that loaded only on the reversed items. In another model, we formed a random intercept factor RIIFA according to Maydeu-Olivares and Garrido (Garrido et al., 2020; Maydeu-Olivares & Coffman, 2006). There, all items loaded on the RIIFA factor, and the loadings were constrained as equal. The RIIFA factor was again set as orthogonal to theoretic factors, which were allowed to correlate with each other. In the RIIFA model, the reverse-worded items were not reverse-coded.

Model fit was evaluated using absolute fit indices such as X^2 and the standardized root mean square residual (SRMR), a parsimony-corrected index called root mean square error of approximation (RMSEA), and comparative fit indices, such as comparative fit index (CFI) and Tucker-Lewis Index (TLI). X^2, SRMR, and RMSEA approach zero when the model fit is good. CFI and TLI approach or exceed 1 in the case of a well-fitting model. When we refer to commonly used cut-offs in the results section, we use guidelines from Schreiber (Schreiber et al., 2006). These were a p-value ≤ .01 for X^2, ≥.95 for CFI, ≥.96 for TLI, ≤.06 for RMSEA, and ≤.08 for SRMR. lavaan produces two types of indices: standard and robust. All reported model fit indices refer to robust indices in this study. Models were also evaluated on their parameters including the inspection of residuals. Tables for essential τ-equivalence and factor loadings in the revised models are presented in the Appendix, Tables S3–S6.

Third, with the best fitting models, we tested measurement invariance across gender and time. The measurement invariance was evaluated stepwise as configural, metric/weak, and scalar/strong invariance. Measurement invariance was evaluated using the following criteria: in the weak compared with the configural invariance, the fit should not decrease by more than .01 in CFI, increase by more than .015 in RMSEA, or increase more than .03 in SRMR. In the strong compared with the weak invariance, the fit should not decrease by more than .01 in CFI, increase by more than .015 in RMSEA, or increase more than .01 in SRMR (Chen, 2007).

Fourth, we estimated the reliability of each subscale to maintain comparability with previous studies. We used alpha, ordinal alpha, and omega as reliability estimates for the subscales. We used the cut-off of ≥.8 to indicate acceptable reliability (Raykov & Marcoulides, 2010). Fifth, the use of the sum score method per each SDQ subscale was evaluated by examining the association between sum scores and factor scores of the factor in the CFA associated with that SDQ scale. The association was estimated with the Spearman rank-order correlation coefficient (ρ). Note that the RIIFA factor from the revised model was not considered, as no corresponding SDQ scale exists. Following Vugteveen (Vugteveen et al., 2020), we consider a Spearman ρ > .85 to be supportive of the continued use of sum scores in practice.

Results

Structural Validity

One-dimensionality and Essential τ-equivalence

The emotional problems scale showed a generally acceptable fit in the one-dimensionality test in all groups, although RMSEA was high in three groups of girls. Similarly, the prosocial behavior scale had an acceptable fit. The conduct problems scale had an acceptable fit based on all other indices except for low TLI values among mid-adolescent girls and early adolescent boys at time 2, and one high RMSEA value among mid-adolescent girls at time 1.

Among girls, the peer problems scale had an acceptable fit based on CFI and SRMR, but other indices such as TLI and RMSEA showed more ambiguous results, denoting that the peer problems scale did

not achieve full one-dimensionality. Among early adolescent girls at time 1, RMSEA was above the limit, and for mid-adolescent girls at time 1 and at time 2, TLI and RMSEA were unacceptable. At time 2 with the early adolescent girls' and boys' groups, the peer problems scale had an acceptable fit. Again, among early adolescent boys at time 1 and mid-adolescent boys at time 2, other indices such as TLI and RMSEA showed an unacceptable fit. This means that the one-dimensionality test was not fully acceptable. At time 1 with mid-adolescent boys' group, TLI, RMSEA, and SRMR showed an unacceptable fit. The hyperactivity-inattention scale had an unacceptable fit in all groups.

Judged by all indices, none of the original scales was essentially τ-equivalent. When looking at the indices individually, some scales had a highly acceptable fit. For instance, in five out of eight groups, the prosocial behavior scale could be seen as essentially τ-equivalent (at time 1 with mid-adolescent girls and boys and early adolescent boys, and at time 2 early and mid-adolescent boys). Similarly, at time 2, the emotional problems scale seemed to have a reasonably acceptable fit in early adolescent girls and boys. Fit indices for essential τ-equivalence are reported in the Appendix, Tables S3 and S4.

It is worth mentioning that a major reason for the poor fit for essential τ-equivalence seemed to be the reverse-worded items. In the Difficulties scales, negatively worded items had a nearly equally sized loading on their factor. For instance, in the conduct problems scale, the reverse-worded item "obedient" had a very low loading on the factor and other items had equally sized loadings. In the hyperactivity-inattention and peer problems' scales, the reverse-worded items had loadings of similar size within the scale, and negatively worded items had loadings of similar size within the scale.

The Multi-Factor Structures

The five-factor structure and the four-factor Total Difficulties structure had an unacceptable fit in all groups. This was not surprising, since not all the scales were one-dimensional. Surprisingly though, at time 1 and time 2 with early adolescent girls' groups, RMSEA and SRMR showed a nearly acceptable fit.

In the revised measurement models, we first tried to add a method factor in the five-factor model that was set orthogonal to the Difficulties factors. All reverse-worded items from the Difficulties factors were regressed on the method factor. The theoretical covariance matrix implied by this model was not

positive definite in four groups. Further inspection revealed that positive indefiniteness was caused by extremely high correlations among the prosocial behavior, conduct problems, and hyperactivity-inattention scales. It seemed that after controlling for the method effect in conduct problems and hyperactivity-inattention scales, they were essentially the same construct as the prosocial behavior.

The Difficulties factors with a method factor model showed a reasonably acceptable fit in all early adolescent groups. In most groups, two items ("obedient" and "adults") had a loading below .3. Among mid-adolescents, all fit values were unacceptable or at the limits of the commonly used cut-offs. When two items ("obedient" and "adults") were removed from the model, the early adolescent student groups showed an acceptable model fit. Among the mid-adolescent groups, the reduced model showed an unacceptable fit. A closer inspection of the model indicated that some other measurement model could work better in these groups.

The Difficulties factors with a RIIFA factor and without "obedient" and "adults" resulted in the best fit in all groups, although it highly resembled the fit in the method factor model. Among early adolescent students, the fit judged by all indices was excellent. Among mid-adolescent students, the fit was better with the time 2 group than with the time 1 group. However, the RIIFA model supported the idea that some other measurement model might work better with mid-adolescent students.

Curiously, the revised measurement models without reversed items in the early adolescent groups could also be considered as essentially τ-equivalent. The fit indices for one-dimensionality, five-factor structure, and revised models are reported below in Tables 3, 4, 5, and 6.

[Table 3.]

[Table 4.]

[Table 5.]

[Table 6.]

Measurement Invariance Across Genders and over Time

Since the RIIFA model had the best fit, we tested measurement invariance with the Difficulties factors and RIIFA model where the items "obedient" and "adults" were removed. The comparisons were

among the early adolescents, between genders, and between times. First, separately at time 1 and time 2, we compared the girls' and boys' groups. Then, separately for the girls and boys, we compared groups at time 1 and 2. This made altogether four groups. Measurement invariance between early and mid-adolescents was not tested because of the unsatisfying model fit among the mid-adolescents.

In general, all tests were statistically significant based on the X^2. However, all tests and all levels of measurement invariance showed an acceptable fit based on CFI, RMSEA, and SRMR. The fit indices were of similar size. The results of the model fit for measurement invariance are reported in Table 7.

[Table 7.]

Reliability

The alpha coefficients were the lowest reliability estimates in all groups, and the ordinal alpha estimated the highest reliabilities. The omega coefficients were closer to alpha than the ordered alpha. The conduct problems scale had the lowest omegas in all groups. The peer problems scale had nearly equally low omegas in five groups. The hyperactivity-inattention scale had a low omega in one group. The emotional problems scale had the highest reliability except at time 2 with early adolescent boys, where the prosocial behavior scale had a higher omega coefficient. However, all reliability estimates for the subscales remained unacceptable except for two: the ordinal alpha for the emotional problems scale and the conduct problems scale without "obedient" at time 2 with mid-adolescent boys (.819 and .806, respectively). The reliability estimates are reported in Table 8 for girls and in Table 9 for boys.

[Table 8.]

[Table 9.]

Resemblance Between Subscale Sum Scores and Factor Scores

The Spearman rank correlations between the SDQ scale sum scores and factor scores were calculated for all Total difficulty subscales and for two extra scales. The peer problems sum score without "adults" and the conduct problems sum score without "obedient" were included, because the factor scores were calculated with the best fitting revised SDQ with a RIIFA factor model where "obedient" and "adults"

were removed. In this way, the scale correlations were as close as possible. In general use, however, full sum scores are used regardless of the many modifications added to the tested factor models.

In general, the correlations were surprisingly low. Only the correlation between the hyperactivity-inattention sum score and factor score indicated a high resemblance. The peer problems scale without "adults" showed acceptable correlations. The lowest correlation (.852) indicates only 73 percent shared variance. Correlations in the emotional problems scale were low in three boys' groups. The conduct problems scale without "obedient" had acceptable correlations in two groups, but all correlations remained quite low, however. The peer problems scale had an almost acceptable correlation in one group. The conduct problems scale showed no acceptable correlations, indicating that 35–50 percent of the variation is not shared. In other words, as much as half of the variation in the conduct problems scale can be counted as noise. The Spearman rank correlations are presented in Table 10.

[Table 10.]

Discussion

The aim of this study was to examine analyze comprehensively the psychometric properties of the SDQ by focusing on structural validity, measurement invariance, reliability, and resemblance of the sum and factor scores. For structural validity, we tested subscale one-dimensionality, essential τ-equivalence of factor loadings, and among early adolescents, measurement invariance across genders and time. Furthermore, we examined competing multi-factor structures: the five-factor structure, the four-factor Total Difficulties structure, the revised SDQ structure with a method factor, and the revised SDQ structure with a RIIFA factor. We calculated three different reliability estimates for each subscale. Finally, to investigate resemblance, we calculated Spearman rank correlations for each SDQ subscale between the sum scores and factor scores. The findings provided mixed support for the structural validity of the SDQ, full support for the measurement invariance, no support for the reliability of the SDQ subscales, and little to no support for the sum score and factor score resemblance and therefore for sum score use.

Structural Validity of the SDQ

Our results revealed that the emotional problems and prosocial behavior scales fulfilled the fit criteria for one-dimensionality. The result corresponds with earlier studies (Koskelainen et al., 2001; Ribeiro Santiago et al., 2021; Richter et al., 2011). The conduct problems scale was deemed one-dimensional, although part of the indices indicated a poor fit especially among mid-adolescent girls. The surprisingly acceptable fit of the conduct problems scale was somewhat contradictory to previous studies. Some studies have reported the peer problems scale being confused with the conduct problems scale (Kim et al., 2015; Muris et al., 2004). In our analysis, the model fit for the peer problems scale was ambiguous. The hyperactivity-inattention scale had a systematically poor fit. A closer inspection revealed that the items might function well but consist of two distinct subfactors. Other studies have also considered subfactors without specifying which attributes these might reflect (Van De Looij-Jansen et al., 2011; Vugteveen et al., 2021). An alternative explanation for the multidimensionality in the hyperactivity-inattention scale is the wording effect. Loadings on the three negatively worded items were similar, and likewise loadings on the two positively worded items were similar. It would be worthwhile trying whether the scale would improve substantially if the reverse-worded items were reversed into negatively worded items.

Our results showed that the five SDQ subscales were not essentially τ-equivalent. Nevertheless, the fit indices would have improved substantially if some items were removed from the conduct problems scale, such as “obedient”. Changes in the scales would make them approach essential τ-equivalence. Consequently, essentially τ-equivalent scales would produce more balanced sum scores, as they have equally weighted items.

Considering that the hyperactivity-inattention and peer problems scales were not one-dimensional, it was expected that the original structure did not fit well. Revision of the measurement model revealed an interesting finding of the close association between prosocial behavior, conduct problems, and hyperactivity-inattention scales. We do not know whether, for example, the conduct problems and hyperactivity-inattention are so opposite to prosocial behavior that they could be considered as the opposing poles of the same continuum. This finding is worth investigating further in future studies. The lack

of discrimination between conduct problems, hyperactivity-inattention, and prosocial behavior could perhaps be understood from the perspective of a p-factor (van Bork et al., 2017). All items are somewhat associated, but that does not mean there exists an actual psychological attribute to explain the general association between them.

The revised measurement model of four Difficulties factors and a method factor with two items removed resulted in an acceptable model fit in all early adolescent students' groups. The items "obedient" and "adults" that were omitted in our study have been frequently reported as problematic (Garrido et al., 2020; Giannakopoulos et al., 2009; Van De Looij-Jansen et al., 2011). This finding adds support for revising or removing these items from the questionnaire.

Considering the two types of fixing the method variance, in all groups, the RIIFA model produced a better fit than the method factor model by Campbell and Fiske and Eid. This is in accordance with what Maydeu-Olivares and Coffman (2006) predicted. Additionally, the factor loadings on the RIIFA were lower than any loading on the theoretic factors, contrary to the method factor loadings. Restricting all loadings to be equal is understandable from the viewpoint that the method, in this case positive versus negative wording, influences all items equally.

An important message from the confirmatory factor models emphasizes the need for cautiousness about using the SDQ for assessing the mental health of mid-adolescents. At time 1 for the mid-adolescent groups, the model fit did not reach the level of the early adolescent groups. At time 2 for the mid-adolescent student groups, the model fit was closer to the fit among younger students but remained unacceptable anyhow. It is possible that the assumed dimensions do not correspond to mid-adolescents' mental health problems. The finding of the higher prevalence of mental health problems among mid-adolescents based on the effect size estimates further underscores the need for a valid and reliable measuring instrument.

Another important message from the revised models is the strong support for the method effect in the SDQ. Several other studies have considered the method effects, and this study is in line with these studies' findings (Duinhof et al., 2019; Garrido et al., 2020; Vugteveen et al., 2020). Although not every

psychometric SDQ study has confirmed the need for controlling the method effect (Black et al., 2021; Español-Martín et al., 2021), this should however highlight the importance of carefully revising the SDQ. Previous studies have been hesitant to suggest changing the SDQ because it might complicate comparisons between the original and modified measures (R. Goodman et al., 2007; Van Roy et al., 2008). Despite this, repeated findings on the problematic structural validity of the SDQ should be taken seriously and lead to change. Several studies have discouraged the use of reverse-worded items (Chyung et al., 2018; Suárez-Alvarez et al., 2018; Weijters & Baumgartner, 2012). One could start the change by rewording the positively worded Difficulties items.

The revised model showed measurement invariance between genders and across time among early adolescents except for the significant X^2 test. The positive findings in gender measurement invariance are in line with previous studies (Bøe et al., 2016; Hoofs et al., 2015; Ortuño-Sierra, Chocarro, Fonseca-Pedrero, et al., 2015), whereas they contradict previous findings among Finnish adolescents (Koskelainen et al., 2001). Strong multi-group invariance encourages the comparison of means between genders and with data collected at different time points.

Sum Scores and Reliability

A low sum score and factor score resemblance including the shortened sum scores imply severe restrictions in using the SDQ as a screening measure of mental health problems and psychological distress. Only the hyperactivity-inattention scale showed a consistently high resemblance between the sum score and factor score. This is interesting because the hyperactivity-inattention scale was not one-dimensional in any of the groups. It was predictable that the sum scores without “obedient” and “adults” corresponded to the factor scores more than the original sum scores. Surprisingly, however, even the shortened sum scores did not correlate highly with the factor scores. Especially in the case of the conduct problems scale, a sum score without “obedient” and a factor score had at least 20 percent bias. Unfortunately, the shortened sum scores are not realistic, since they are rarely used in practice. Furthermore, in the original conduct problems sum score, the proportion of bias was as much as half of the variation. Therefore, this study

indicates that especially the conduct problems sum score is not a reliable assessment tool for adolescent mental health.

Several studies have recommended the use of the Total Difficulties sum score over the subscale sum scores. This, however, does not exclusively solve the problem of unreliability or invalidity. The Total Difficulties sum score is nothing more than a sum of the subscale scores, and when three of the four subscales may misclassify a great deal of the respondents, one should desist from using the sum score in its current form.

Alpha coefficients should only be considered in comparison to previous studies and other estimates of reliability. They should not be considered as reliability coefficients of the subscales, because not all scales were essentially τ-equivalent, let alone one-dimensional. Therefore, alpha is a biased reliability estimate. Omega coefficients are considered less biased in case of misfit. All omega coefficients were low, and they were especially low on the conduct problems scale. Let us not forget that “obedient” had little in common with other items. The reliability of the conduct problems scale has been reported as low in numerous previous studies, as has the reliability of the peer problems scale (Muris et al., 2003; Rønning et al., 2004; Van Roy et al., 2008).

Our study showed, however, that even the emotional problems and prosocial behavior scales had low omega coefficients. When sum scores are used to make decisions on the individual level, an internal consistency reliability of at least .80 or .85 is required for “lower-stakes standardized tests,” while “high-stakes standardized tests” should have a reliability as high as .90 (Wells & Wollack, 2003). Low reliability estimates combined with the findings of the sum score and factor score resemblance should have consequences for the use of the SDQ in screening and research. In screening, as many as half of the true positives can be misclassified and adolescents might remain without the treatment they need.

In research, these findings should encourage researchers to use latent variable modeling to account for the measurement error when one-fifth to one-half of the variation in the scale is considered as measurement error. Reliability is especially important when studies use correlations or covariances. A low reliability lowers group correlations according to Spearman’s attenuation formula. In a multivariate

regression model, a low reliability lowers the explained variation, and it has unpredictable effects on the regression coefficients. Thus, a sum score used as an explanatory or explained variable may produce biased results.

Strengths and Limitations

The findings in our study contribute to the long and lively discussion of the structural validity and reliability of the SDQ scores. There are several strengths that must be acknowledged. First, this study aimed to respond to several recently published studies investigating the SDQ by using similar methods. We modelled the method effects with two different types of method factors, in line with several previous studies (Garrido et al., 2020; Hoofs et al., 2015; Van De Looij-Jansen et al., 2011; Van Roy et al., 2008; Vugteveen et al., 2020, 2021). We also estimated reliability with the most frequently used reliability estimates to maintain comparability with most of the previously published SDQ studies. Finally, we estimated the resemblance between SDQ sum scores and factor scores (Vugteveen et al., 2020). These results enable research on the SDQ to accumulate and researchers and clinicians to make evidence-based, carefully considered decisions regarding whether to use the SDQ and how to use it.

Second, this study provides solid support for the method effect in the SDQ self-report. Since the preliminary psychometric study on the SDQ by Goodman (R. Goodman, 2001), researchers have reflected on the possible method-related variance in the SDQ responses. It is possible that the method effect varies across countries, and perhaps not all language versions of the SDQ need to be changed. Changing the wording in only some languages could complicate cross-country comparisons, however.

Third, no previous study has tested the structural validity of the SDQ sum score, even though it is the one measurement model used in studies. It is important to carefully inspect the validity of the constructs according to their use, and not only to conduct a routine check with traditional alpha coefficients. Psychological assessment has been criticized for a lack of rigor in measurement quality and reporting, and we must start to pay more attention to this (Jessica K. Flake et al., 2017; Jessica Kay Flake & Fried, 2020).

Fourth, this study covered altogether eight nationally representative groups of adolescents: girls and boys at different stages of adolescence and from different cohorts. Several groups enabled us to compare findings between genders, age groups, and over time. We could detect that the SDQ seemed to function better among the early adolescents compared to the mid-adolescents. In addition, we could see that after revising the model, the gender differences disappeared, and multi-group measurement invariance was achieved.

No study is without limitations. First, we are aware that time has passed since the data were collected. However, we chose the data because of their high quality, and furthermore no newer data on Finnish adolescents were available. Second, obviously, cross-sectional data do not allow for real causal inferences, let alone statements on within-person effects. Cross-sectional studies are nevertheless suitable for psychometric studies, and they enable explorations for signs of potential causal relationships that require further longitudinal or even intensive ecological momentary assessment investigations.

Conclusions

This study supports the revision of the SDQ. The questionnaire has very good elements that seem to be stable over time, but some parts of it need to be updated. Certain items, such as “obedient” and “adults,” could be removed or carefully reworded. The current length of the questionnaire has worked well for respondents and researchers, so rewording or changing items could be prioritized. The rewording would warrant a new study where, first, multiple groups of practitioners and adolescents are interviewed to reach an up-to-date understanding of suitable items. Second, a large-scale psychometric study should be conducted to investigate the structural validity of the developed questionnaire and to drop the worst functioning items. Similar questionnaire development has been done in quality-of-life research (Skevington et al., 2004), for example.

For researchers, the revised SDQ Total Difficulties measurement model with a RIIFA factor seems to perform well and enable mean comparisons. However, because of the consistent findings of low reliability in the original SDQ subscales, a large-scale psychometric study should be conducted to examine the

structural validity and reliability of the new, developed questionnaire. Until then, practitioners should desist from using the SDQ in its current form. When the new psychometrically sound questionnaire is available, practitioners should then be widely informed about the suggested changes to gain more reliable use of the SDQ. The items were chosen decades ago, and today's children and adolescents may interpret the questions differently. Goodman (R. Goodman, 1997) wrote about the Rutter questionnaires: they have generally worn well, though they do show their age in some ways. It is time to give honor to Goodman's great work and start actively revising the SDQ.

References

Ahmad Ghanizadeh, Ahad Izadpanah, & Gholamreza Abdollahi. (2007). Scale Validation of the Strengths and Difficulties Questionnaire in Iranian Children [Article]. *Iranian Journal of Psychiatry*, *2*(2).

Barkmann, C., & Schulte-Markwort, M. (2012). Prevalence of emotional and behavioural disorders in German children and adolescents: A meta-analysis. *Journal of Epidemiology and Community Health*, *66*(3), 194–203. https://doi.org/10.1136/jech.2009.102467

Becker, A., Wang, B., Kunze, B., Otto, C., Schlack, R., Hölling, H., Ravens-Sieberer, U., Klasen, F., Rogge, J., Isensee, C., & Rothenberger, A. (2018). Normative data of the self-report version of the German strengths and difficulties questionnaire in an epidemiological setting. *Zeitschrift Fur Kinder- Und Jugendpsychiatrie Und Psychotherapie*, *46*(6), 523–533. https://doi.org/10.1024/1422-4917/a000589

Black, L., Mansfield, R., & Panayiotou, M. (2021). Age Appropriateness of the Self-Report Strengths and Difficulties Questionnaire. *Assessment*, *28*(6), 1556–1569. https://doi.org/10.1177/1073191120903382

Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): Factor structure and gender equivalence in Norwegian adolescents. *PLoS ONE*, *11*(5), 1–16. https://doi.org/10.1371/journal.pone.0152202

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

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*(2), 79–88. https://doi.org/10.1027/1015-5759.23.2.79

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and the clinical implications of low reliability. *Journal of General Psychology*, *130*(3), 290–304. https://doi.org/10.1080/00221300309601160

Charter, R. A., & Feldt, L. S. (2001). Meaning of reliability in terms of correct and incorrect clinical decisions: The art of decision making is still alive. *Journal of Clinical and Experimental Neuropsychology*, *23*(4), 530–537. https://doi.org/10.1076/jcen.23.4.530.1227

Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Chyung, S. Y. Y., Barkin, J. R., & Shamsy, J. A. (2018). Evidence-Based Survey Design: The Use of Negatively Worded Items in Surveys. *Performance Improvement*, *57*(3), 16–25. https://doi.org/10.1002/pfi.21749

Clayborne, Z. M., Varin, M., & Colman, I. (2019). Systematic Review and Meta-Analysis: Adolescent Depression and Long-Term Psychosocial Outcomes. *Journal of the American Academy of Child and Adolescent Psychiatry*, *58*(1), 72–79. https://doi.org/10.1016/j.jaac.2018.07.896

Davidov, M., Vaish, A., Knafo-Noam, A., & Hastings, P. D. (2016). The Motivational Foundations of Prosocial Behavior From A Developmental Perspective–Evolutionary Roots and Key Psychological Mechanisms: Introduction to the Special Section. *Child Development*, *87*(6), 1655–1667. https://doi.org/10.1111/cdev.12639

De Vries, P. J., Davids, E. L., Mathews, C., & Aarø, L. E. (2018). Measuring adolescent mental health around the globe: Psychometric properties of the self-report Strengths and Difficulties Questionnaire in South Africa, and comparison with UK, Australian and Chinese data. *Epidemiology and Psychiatric Sciences*, *27*(4), 369–380. https://doi.org/10.1017/S2045796016001207

Duinhof, E. L., Lek, K. M., De Looze, M. E., Cosma, A., Mazur, J., Gobina, I., Wüstner, A., Vollebergh, W. A. M., & Stevens, G. W. J. M. (2019). Revising the self-report strengths and difficulties questionnaire for

cross-country comparisons of adolescent mental health problems: The SDQ-R. *Epidemiology and Psychiatric Sciences*. https://doi.org/10.1017/S2045796019000246

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*, 1–15. https://doi.org/10.1186/1753-2000-2-8

Eid, M. (2000). A multitrait-multimethod model with minimal assumptions. *Psychometrika*, *65*(2), 241–261. https://doi.org/https://doi.org/10.1007/BF02294377

Ellonen, N., Fagerlund, M., Kääriäinen, J., Peltola, M., & Sariola, H. (2013). *Child Victim Survey 2013.* Finnish Social Science Data Archive [distributor].

Ellonen, N., Kääriäinen, J., Salmi, V., & Sariola, H. (2008). *Child Victim Survey 2008*. Finnish Social Science Data Archive [distributor].

Español-Martín, G., Pagerols, M., Prat, R., Rivas, C., Sixto, L., Valero, S., Artigas, M. S., Ribasés, M., Ramos-Quiroga, J. A., Casas, M., & Bosch, R. (2021). Strengths and Difficulties Questionnaire: Psychometric Properties and Normative Data for Spanish 5- to 17-Year-Olds. *Assessment*, *28*(5), 1445–1458. https://doi.org/10.1177/1073191120918929

Essau, C. A., Olaya, B., Anastassiou-Hadjicharalambous, X., Pauli, G., Gilvarry, C., Bray, D., O'callaghan, J., & Ollendick, T. H. (2012). Psychometric properties of the Strength and Difficulties Questionnaire from five European countries. *International Journal of Methods in Psychiatric Research*, *21*(3), 232–245. https://doi.org/https://doi.org/10.1002/mpr.1364

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, *3*(4), 456–465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct Validation in Social and Personality Research: Current Practice and Recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378. https://doi.org/10.1177/1948550617693063

Garrido, L. E., Barrada, J. R., Aguasvivas, J. A., Martínez-Molina, A., Arias, V. B., Golino, H. F., Legaz, E., Ferrís, G., & Rojo-Moreno, L. (2020). Is Small Still Beautiful for the Strengths and Difficulties Questionnaire? Novel Findings Using Exploratory Structural Equation Modeling. *Assessment*, *27*(6), 1349–1367. https://doi.org/10.1177/1073191118780461

Giannakopoulos, G., Tzavara, C., Dimitrakaki, C., Kolaitis, G., Rotsika, V., & Tountas, Y. (2009). The factor structure of the Strengths and Difficulties Questionnaire (SDQ) in Greek adolescents. *Annals of General Psychiatry*, *8*, 20. https://doi.org/10.1186/1744-859X-8-20

Gomez, R., Motti-Stefanidi, F., Jordan, S., & Stavropoulos, V. (2021). Greek Validation of the Factor Structure and Longitudinal Measurement Invariance of the Strengths and Difficulties Questionnaire-Self Report (SDQ-SR): Exploratory Structural Equation Modelling. *Child Psychiatry and Human Development*, *52*(5), 880–890. https://doi.org/10.1007/s10578-020-01065-7

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the strengths and difficulties questionnaire (SDQ): Data from british parents, teachers, and children. *Journal of Abnormal Child Psychology*, *38*(8), 1179–1191. https://doi.org/10.1007/s10802-010-9434-x

Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *38*(5), 581–586. https://doi.org/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 and Adolescent Psychiatry*, *40*(11), 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

Goodman, R., Iervolino, A. C., Collishaw, S., Pickles, A., & Maughan, B. (2007). Seemingly minor changes to a questionnaire can make a big difference to mean scores: A cautionary tale. *Social Psychiatry and Psychiatric Epidemiology*, *42*(4), 322–327. https://doi.org/10.1007/s00127-007-0169-0

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*(4), 1–21. https://doi.org/10.1371/journal.pone.0120930

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling*.

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 Investig*, *12*(4), 491–499. https://doi.org/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 and Adolescent Psychiatry*, *10*(3), 180–185. https://doi.org/10.1007/s007870170024

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*(6). https://doi.org/10.1016/j.comppsych.2013.01.002

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–362. https://doi.org/10.1037/1082-989X.11.4.344

Mellor, D., & Stokes, M. (2007). The factor structure of the strengths and difficulties questionnaire. *European Journal of Psychological Assessment*, *23*(2), 105–112. https://doi.org/10.1027/1015-5759.23.2.105

Merikukka, M., Ristikari, T., Tuulio-Henriksson, A., Gissler, M., & Laaksonen, M. (2018). Childhood determinants for early psychiatric disability pension: A 10-year follow-up study of the 1987 Finnish Birth Cohort. *International Journal of Social Psychiatry*, *64*(8), 715–725. https://doi.org/10.1177/0020764018806936

Muris, P., Meesters, C., Eijkelenboom, A., & Vincken, M. (2004). The self-report version of the Strengths and Difficulties Questionnaire: Its psychometric properties in 8- to 13-year-old non-clinical children. *The British Journal of Clinical Psychology*, *43*(4), 437–448. https://doi.org/https://doi.org/10.1348/0144665042388982

Muris, P., Meesters, C., & Van den Berg, F. (2003). The Strengths and Difficulties Questionnaire (SDQ) further evidence for its reliability and validity in a community sample of Dutch children and adolescents. *European Child and Adolescent Psychiatry*, *12*(1), 1–8. https://doi.org/10.1007/s00787-003-0298-2

Nieto, M. D., Garrido, L. E., Martínez-Molina, A., & Abad, F. J. (2021). Modeling Wording Effects Does Not Help in Recovering Uncontaminated Person Scores: A Systematic Evaluation with Random Intercept Item Factor Analysis. *Frontiers in Psychology*, *12*(June), 1–24. https://doi.org/10.3389/fpsyg.2021.685326

Ortuño-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., Riba, S. S. I., & Muñiz, J. (2015). The assessment of emotional and Behavioural problems: Internal structure of The Strengths and Difficulties Questionnaire. *International Journal of Clinical and Health Psychology*, *15*(3), 265–273. https://doi.org/10.1016/j.ijchp.2015.05.005

Ortuño-Sierra, J., Fonseca-Pedrero, E., Aritio-Solana, R., Velasco, A. M., de Luis, E. C., Schumann, G., Cattrell, A., Flor, H., Nees, F., Banaschewski, T., Bokde, A., Whelan, R., Buechel, C., Bromberg, U., Conrod, P., Frouin, V., Papadopoulos, D., Gallinat, J., Garavan, H., … Lawrence, C. (2015). New evidence of factor structure and measurement invariance of the SDQ across five European nations. *European Child and Adolescent Psychiatry*, *24*(12), 1523–1534. https://doi.org/10.1007/s00787-015-0729-x

Ortuño-Sierra, J., Fonseca-Pedrero, E., Paino, M., Sastre I Riba, S., & Muñiz, J. (2015). Screening mental health problems during adolescence: Psychometric properties of the Spanish version of the strengths and difficulties Questionnaire. *Journal of Adolescence*, *38*, 49–56. https://doi.org/10.1016/j.adolescence.2014.11.001

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*(1), 43–48. https://doi.org/10.1027/1015-5759.24.1.43

Philipp, J., Zeiler, M., Waldherr, K., Truttmann, S., Dür, W., Karwautz, A. F. K., & Wagner, G. (2018). Prevalence of emotional and behavioral problems and subthreshold psychiatric disorders in Austrian

adolescents and the need for prevention. *Social Psychiatry and Psychiatric Epidemiology*, *53*(12), 1325–1337. https://doi.org/10.1007/s00127-018-1586-y

Raykov, T., & Marcoulides, G. A. (2010). *Introduction to psychometric theory* [Book]. Routledge. https://doi.org/https://doi.org/10.4324/9780203841624

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Ribeiro Santiago, P. H., Manzini, D., Haag, D., Roberts, R., Smithers, L. G., & Jamieson, L. (2021). Exploratory Graph Analysis of the Strengths and Difficulties Questionnaire in the Longitudinal Study of Australian Children. *Assessment*. https://doi.org/10.1177/10731911211024338

Richter, J., Sagatun, Å., Heyerdahl, S., Oppedal, B., & Røysamb, E. (2011). The Strengths and Difficulties Questionnaire (SDQ) - Self-Report. An analysis of its structure in a multiethnic urban adolescent sample. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *52*(9), 1002–1011. https://doi.org/10.1111/j.1469-7610.2011.02372.x

Rønning, J. A., Handegaard, B. H., Sourander, A., & Mørch, W. T. (2004). The Strengths and Difficulties Self-Report Questionnaire as a screening instrument in Norwegian community samples. *European Child and Adolescent Psychiatry*, *13*(2), 73–82. https://doi.org/10.1007/s00787-004-0356-4

Rosseel, Y. (2012). lavaan: an R package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Ruchkin, V., Koposov, R., & Schwab-Stone, M. (2007). The strength and difficulties questionnaire: Scale validation with Russian adolescents. *Journal of Clinical Psychology*, *63*(9), 861–869. https://doi.org/https://doi.org/10.1002/jclp.20401

Santiago, P. H. R., Manzini Macedo, D., Haag, D., Roberts, R., Smithers, L., Hedges, J., & Jamieson, L. (2021). Exploratory Graph Analysis of the Strengths and Difficulties Questionnaire for Aboriginal and/or Torres Strait Islander Children. *Frontiers in Psychology*, *12*(August), 1–20. https://doi.org/10.3389/fpsyg.2021.573825

Schreiber, J. B., Stage, F. K., King, J., Nora, A., & Barlow, E. A. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *Journal of Educational Research*, *99*(6), 323–338. https://doi.org/10.3200/JOER.99.6.323-338

Skevington, S. M., Sartorius, N., Amir, M., Sartorius, N., Orley, J., Kuyken, W., Power, M., Herrman, H., Schofield, H., Murphy, B., Metelko, Z., Szabo, S., Pibernik-Okanovic, M., Quemada, N., Caria, A., Rajkumar, S., Kumar, S., Saxena, S., Baron, D., … van Dam, F. (2004). Developing methods for assessing quality of life in different cultural settings - The history of the WHOQOL instruments. *Social Psychiatry and Psychiatric Epidemiology*, *39*(1), 1–8. https://doi.org/10.1007/s00127-004-0700-5

Stevanovic, D., Urbán, R., Atilola, O., Vostanis, P., Singh Balhara, Y. P., Avicenna, M., Kandemir, H., Knez, R., Franic, T., & 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. https://doi.org/10.1017/S2045796014000201

Suárez-Alvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in likert scales: A questionable practice. *Psicothema*, *30*(2), 149–158. https://doi.org/10.7334/psicothema2018.33

van Bork, R., Epskamp, S., Rhemtulla, M., Borsboom, D., & van der Maas, H. L. J. (2017). What is the p-factor of psychopathology? Some risks of general factor modeling. *Theory and Psychology*, *27*(6), 759–773. https://doi.org/10.1177/0959354317737185

Van De Looij-Jansen, P. M., Goedhart, A. W., De Wilde, E. J., & Treffers, P. D. A. (2011). Confirmatory factor analysis and factorial invariance analysis of the adolescent self-report Strengths and Difficulties Questionnaire: How important are method effects and minor factors? *British Journal of Clinical Psychology*, *50*(2), 127–144. https://doi.org/10.1348/014466510X498174

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, and Allied Disciplines, *49*(12), 1304–1312. https://doi.org/10.1111/j.1469-7610.2008.01942.x

Vugteveen, J., de Bildt, A., Serra, M., de Wolff, M. S., & Timmerman, M. E. (2020). Psychometric Properties of the Dutch Strengths and Difficulties Questionnaire (SDQ) in Adolescent Community and Clinical Populations. *Assessment*, *27*(7), 1476–1489. https://doi.org/10.1177/1073191118804082

Vugteveen, J., de Bildt, A., Theunissen, M., Reijneveld, M., & Timmerman, M. (2021). Validity Aspects of the Strengths and Difficulties Questionnaire (SDQ) Adolescent Self-Report and Parent-Report Versions Among Dutch Adolescents. *Assessment*, *28*(2), 601–616. https://doi.org/10.1177/1073191119858416

Weijters, B., & Baumgartner, H. (2012). Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research*, *49*(5), 737–747. https://doi.org/10.1509/jmr.11.0368

Weir, K., & Duveen, G. (1981). Further Development and Validation of the Prosocial Behaviour Questionnaire for Use by Teachers. *Journal of Child Psychology and Psychiatry*, *22*(4), 357–374. https://doi.org/10.1111/j.1469-7610.1981.tb00561.x

Wells, C. S., & Wollack, J. a. (2003). An Instructor's Guide to Understanding Test Reliability. *Testing and Evaluation Services*, 2–5.

Yao, S., Zhang, C., Zhu, X., Jing, X., McWhinnie, C. M., & Abela, J. R. Z. (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*(1), 55–62. https://doi.org/10.1016/j.jadohealth.2008.11.006

Table 1. *Background information of the previous studies cited in this paper.*

Study #	Author	Country	Version	Participants' age	Sample size
1	(Ribeiro Santiago et al., 2021)	Australia	caregiver	4 to 10	20,000
2	(Santiago et al., 2021)	Australia	caregiver	4 to 10	4,000
3	(Black et al., 2021)	England	self	11 to 15	30,290
4	(Español-Martín et al., 2021)	Spain	self	5 to 17	2,018
5	(Garrido et al., 2020)	Spain	self	10 to 18	67,253
6	(Duinhof et al., 2019)	Bulgaria, Germany, Greece, Netherlands, Poland, Romania, Slovenia	self	11, 13, 15	33,233
7	(Gomez et al., 2021)	Greece	self	12 to 17.9	968
8	(Vugteveen et al., 2020)	Netherlands	self	12 to 17	5,081
9	(Vugteveen et al., 2021)	Netherlands	self	12 to 17	4,053
10	(Becker et al., 2018)	Germany	self	11 to 17	6,726
11	(De Vries et al., 2018)	South Africa	self	13	3,451
12	(Bøe et al., 2016)	Norway	self	16 to 18	10,254
13	(Ortuño-Sierra, Fonseca-Pedrero, Aritio-Solana, et al., 2015)	Spain, England, Ireland, Germany, France	self	12 to 17	3,012
14	(Ortuño-Sierra, Fonseca-Pedrero, Paino, et al., 2015)	Spain	self	14 to 18	1,474
15	(Ortuño-Sierra, Chocarro, Fonseca-Pedrero, et al., 2015)	Spain	self	11 to 19	1,547
16	(Stevanovic et al., 2015)	India, Indonesia, Nigeria, Serbia, Turkey, Bulgaria, Croatia	self	13 to 18	2,367
17	(Hoofs et al., 2015)	Netherlands	self	mean = 14.07	11,207
18	(Kim et al., 2015)	Republic of Korea	self	11 to 16	3,199
19	(Liu et al., 2013)	Taiwan	self	6 to 15	3,899
20	(Essau et al., 2012)	Germany, Cyprus, England, Sweden, Italy	self	12 to 17	2,418
21	(Van De Looij-Jansen et al., 2011)	Netherlands	self	11 to 16	12,795
22	(Richter et al., 2011)	Norway	self	15 to 16	7,343
23	(A. Goodman et al., 2010)	UK (England)	self	11 to 16	7,678
24	(Giannakopoulos et al., 2009)	Greece	self	11 to 17	1,914
25	(Yao et al., 2009)	China	self	11 to 18	1,135
26	(Du et al., 2008)	China	self	11 to 17	816
27	(Van Roy et al., 2008)	Norway	self	10 to 19	26,269
28	(Percy et al., 2008)	Ireland	self	12	3,753
29	(Ruchkin et al., 2007)	Russia	self	13 to 18	2,892
30	(Capron et al., 2007)	France	self	mean = 12.8	1,400
31	(Mellor & Stokes, 2007)	Australia	self	7 to 17	914
32	(Ahmad Ghanizadeh et al., 2007)	Iran	self	3 to 18	756
33	(Rønning et al., 2004)	Norway	self	11 to 16	4,167
34	(Muris et al., 2004)	Netherlands	self	8 to 13	439
35	(Muris et al., 2003)	Netherlands	self	mean = 12.3	562
36	(Koskelainen et al., 2001)	Finland	self	13 to 17	1,458
37	(R. Goodman, 2001)	UK (England)	self	5 to 15	3,983

Table 2. *The results of the reviewed studies.*

	Theme	Studies
Structural validity	Investigated structural validity or dimensionality	All 37 studies except [32]
	Support for the five factors	[4], [10], [22], [24], [25], [27], [29], [30], [32], [35], [36], [37]
	Support for the number of factors:	3 factors: [7], [20], [23]; 4 factors: [34]; 6 factors: [8], [9]
	Against the five-factor structure	[2], [3], [5], [7], [8], [9], [11], [12], [13], [14], [15], [16], [17], [18], [21], [26], [28], [31], [33]
	Problematic item: "Obedient"	[1], [2], [5], [7], [10], [17], [18], [19], [20], [22], [24], [26], [27], [28], [29], [30], [35], [36], [37]
	Problematic item: "Adults"	[1], [2], [5], [6], [13], [15], [21], [22], [23], [24], [30], [32], [36]
	Problematic item: "Persistent"	[1], [2], [5], [6], [7], [8], [10], [11], [13], [14], [16], [19], [20], [22], [24], [26], [28], [36], [37]
	Used a method factor	[5], [8], [9], [17], [21], [27]
	Reverse-worded items loading on Prosocial behavior factor	[5], [6], [8], [9], [11], [13], [14], [17], [20], [21], [24], [27], [28], [34], [36], [37]
	Tested measurement invariance	[3], [4], [5], [6], [7], [8], [12], [13], [14], [15], [16], [17], [20], [21], [22], [25]
	Measurement invariance across genders	[4], [5], [12], [14], [15], [17], [21], [25]
	Measurement invariance across age	[3], [5], [14], [15], [21]
Reliability	Investigated reliability	All, except [16], [28]
	Used Alpha	[3], [4], [7], [8], [9], [10], [11], [12], [14], [17], [18], [19], [20], [21], [23], [24], [25], [26], [27], [29], [30], [31], [32], [33], [34], [35], [36], [37]
	Used Ordinal alpha	[3], [4], [5], [6], [12], [13], [15], [21]
	Used Omega	[1], [2], [3], [4], [7], [8], [10]
	Used Other reliability estimate	[5], [9], [12], [17], [18], [19], [20], [22], [25], [30], [35], [37]
	Low reliability in Peer Problems	[1], [4], [5], [6], [9], [11], [13], [14], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [29], [30], [32], [33], [34], [35], [36], [37]
	Low reliability in Conduct Problems	[1], [4], [5], [7], [9], [10], [11], [13], [14], [17], [18], [19], [20], [21], [22], [23], [24], [26], [27], [29], [30], [33], [34], [35], [36], [37]
	Low reliability in Prosocial Behavior	[5], [9], [11], [17], [18], [21], [22], [29], [30], [34], [35]
	Low reliability in Hyperactivity	[5], [11], [19], [21], [22], [29]
	Low reliability in Emotional Problems	[5], [11], [26], [34]
Support for the use of scale scores	For the Total score and subscales	[4], [8], [9], [10], [11], [12], [23], [25], [30], [32], [36], [37]
	For the Total score only	[17], [18], [20], [22], [28], [34], [35]
	Against scoring	[1], [2], [3], [5], [7], [13], [16], [24], [26], [27], [29], [31], [33]

Table 3. *Confirmatory factor models in early adolescent girls' groups.*

Girls	2008: early adolescent	obs	X^2	df	p	CFI	TLI	RMSEA	SRMR
1	EPs	3,679	100.886	5	<.001	.983	.966	.072	.037
2	PPs	3,656	83.594	5	<.001	.954	.908	.066	.047
3	CPs	3,697	15.128	5	.010	.993	.987	.023	.023
4	HA	3,660	303.965	5	<.001	.896	.793	.128	.071
5	PS	3,702	33.056	5	<.001	.992	.985	.039	.023
6	5 (EPs, PPs, CPs, HA, and PS) factors	3,459	3520.799	265	<.001	.861	.842	.060	.072
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	3,491	1851.044	164	<.001	.908	.893	.054	.061
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	3,491	1338.565	159	<.001	.936	.923	.046	.054
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	3,503	974.697	125	<.001	.952	.941	.044	.049
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	3,503	923.907	128	<.001	.955	.946	.042	.047
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	3,592	982.662	62	<.001	.929	.924	.064	.070
	2013: early adolescent								
1	EPs	2,932	31.761	5	<.001	.992	.984	.043	.024
2	PPs	2,920	41.558	5	<.001	.969	.938	.050	.040
3	CPs	2,928	3.774	5	.582	1	1.003	.000	.014
4	HA	2,898	205.322	5	<.001	.904	.809	.118	.065
5	PS	2,929	19.648	5	.001	.994	.989	.032	.021
6	5 (EPs, PPs, CPs, HA, and PS) factors	2,729	2144.974	265	<.001	.880	.864	.051	.064
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	2,770	1056.358	164	<.001	.927	.915	.044	.052
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	2,770	643.819	159	<.001	.960	.952	.033	.043
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	2,788	526.049	125	<.001	.965	.958	.034	.042
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	2,788	558.220	128	<.001	.963	.956	.035	.043
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	2,865	477.627	62	<.001	.949	.945	.048	.060

EPs = Emotional problems, PPs = Peer problems, CPs = Conduct problems, HA = Hyperactivity, PS = Prosocial behavior, obs = observations, df = degrees of freedom, RIIFA = Random Intercept Item Factor Analysis, item 7 = "Obedient," item 23 = "Adults"

Table 4. *Confirmatory factor models in mid-adolescent girls' groups.*

Girls	2008: mid-adolescent	obs	X^2	df	p	CFI	TLI	RMSEA	SRMR
1	EPs	2,765	68.927	5	<.001	.986	.972	.068	.034
2	PPs	2,755	103.883	5	<.001	.941	.881	.085	.060
3	CPs	2,761	68.318	5	<.001	.950	.901	.068	.053
4	HA	2,752	247.858	5	<.001	.932	.864	.133	.067
5	PS	2,761	31.036	5	<.001	.990	.979	.043	.027
6	5 (EPs, PPs, CPs, HA, and PS) factors	2,656	3461.914	265	<.001	.819	.795	.067	.082
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	2,676	2117.443	164	<.001	.863	.842	.067	.076
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	2,676	1801.196	159	<.001	.885	.863	.062	.071
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	2,684	1334.682	125	<.001	.911	.891	.060	.066
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	2,684	1309.820	128	<.001	.913	.896	.059	.065
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	2,732	1105.278	62	<.001	.896	.889	.078	.086
	2013: mid-adolescent								
1	EPs	2,499	94.857	5	<.001	.979	.957	.085	.041
2	PPs	2,498	101.731	5	<.001	.949	.898	.088	.058
3	CPs	2,497	31.308	5	<.001	.975	.950	.046	.037
4	HA	2,486	247.874	5	<.001	.947	.895	.140	.068
5	PS	2,492	47.741	5	<.001	.986	.972	.059	.032
6	5 (EPs, PPs, CPs, HA, and PS) factors	2,415	3036.555	265	<.001	.851	.832	.066	.082
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	2,433	1796.482	164	<.001	.891	.874	.064	.072
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	2,433	1419.556	159	<.001	.916	.899	.057	.066
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	2,440	1042.173	125	<.001	.936	.921	.055	.061
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	2,440	1021.025	128	<.001	.937	.925	.053	.059
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	2,472	958.811	62	<.001	.911	.906	.077	.085

EPs = Emotional problems, PPs = Peer problems, CPs = Conduct problems, HA = Hyperactivity, PS = Prosocial behavior, obs = observations, df = degrees of freedom, RIIFA = Random Intercept Item Factor Analysis, item 7 = "Obedient," item 23 = "Adults"

Table 5. *Confirmatory factor models in early adolescent boys' groups.*

boys	2008: early adolescent	obs	X^2	df	p	CFI	TLI	RMSEA	SRMR
1	EPs	3,556	50.352	5	<.001	.986	.973	.051	.031
2	PPs	3,543	132.150	5	<.001	.928	.857	.085	.058
3	CPs	3,559	16.482	5	.006	.992	.983	.025	.025
4	HA	3,522	311.504	5	<.001	.866	.732	.132	.076
5	PS	3,585	30.233	5	<.001	.994	.988	.038	.020
6	5 (EPs, PPs, CPs, HA, and PS) factors	3,323	4664.316	265	<.001	.776	.746	.071	.090
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	3,351	2163.133	164	<.001	.862	.840	.060	.070
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	3,351	1350.729	159	<.001	.918	.902	.047	.058
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	3,375	887.124	125	<.001	.945	.933	.043	.051
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	3,375	794.892	128	<.001	.952	.942	.039	.049
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	3,468	653.763	62	<.001	.939	.935	.052	.067
					<.001				
	2013: early adolescent								
1	EPs	2,876	26.852	5	<.001	.991	.982	.039	.026
2	PPs	2,868	55.166	5	<.001	.971	.941	.059	.041
3	CPs	2,875	38.698	5	<.001	.973	.946	.048	.040
4	HA	2,827	163.091	5	<.001	.925	.851	.106	.059
5	PS	2,900	45.169	5	<.001	.988	.977	.053	.028
6	5 (EPs, PPs, CPs, HA, and PS) factors	2,614	3166.588	265	<.001	.816	.791	.065	.083
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	2,652	1402.915	164	<.001	.891	.874	.053	.063
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	2,652	812.239	159	<.001	.943	.931	.039	.050
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	2,678	599.992	125	<.001	.956	.946	.038	.046
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	2,678	559.738	128	<.001	.960	.952	.035	.045
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	2,771	424.952	62	<.001	.949	.946	.046	.064

EPs = Emotional problems, PPs = Peer problems, CPs = Conduct problems, HA = Hyperactivity, PS = Prosocial behavior, obs = observations, df = degrees of freedom, RIIFA = Random Intercept Item Factor Analysis, item 7 = "Obedient," item 23 = "Adults"

Table 6. *Confirmatory factor models in mid-adolescent boys' groups.*

boys	2008: mid-adolescent	obs	X^2	df	p	CFI	TLI	RMSEA	SRMR
1	EPs	2,726	53.787	5	<.001	.983	.967	.060	.039
2	PPs	2,736	163.747	5	<.001	.914	.829	.108	.075
3	CPs	2,731	30.235	5	<.001	.981	.961	.043	.032
4	HA	2,726	312.422	5	<.001	.889	.777	.150	.081
5	PS	2,740	21.866	5	.001	.994	.989	.035	.020
6	5 (EPs, PPs, CPs, HA, and PS) factors	2,602	4279.769	265	<.001	.744	.711	.076	.098
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	2,622	2325.548	164	<.001	.820	.791	.071	.083
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	2,622	1662.068	159	<.001	.875	.850	.060	.070
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	2,633	1126.367	125	<.001	.912	.892	.055	.063
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	2,633	1081.241	128	<.001	.916	.900	.053	.062
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	2,667	847.509	62	<.001	.901	.894	.069	.085
	2013: mid-adolescent								
1	EPs	2,283	26.002	5	<.001	.998	.987	.043	.026
2	PPs	2,294	142.127	5	<.001	.940	.879	.109	.069
3	CPs	2,293	7.027	5	.219	.999	.998	.013	.015
4	HA	2,301	442.801	5	<.001	.871	.741	.195	.099
5	PS	2,290	34.610	5	<.001	.988	.975	.051	.027
6	5 (EPs, PPs, CPs, HA, and PS) factors	2,169	4168.439	265	<.001	.773	.743	.082	.102
7	Difficulties (EPs, PPs, CPs, HA) factors (sum score model)	2,202	2383.272	164	<.001	.844	.819	.078	.081
8	Difficulties (EPs, PPs, CPs, HA) factors and a method factor	2,202	1399.432	159	<.001	.913	.896	.060	.063
9	Difficulties (EPs, PPs, CPs, HA) factors and a method factor without items 7 and 23	2,210	990.930	125	<.001	.936	.922	.056	.057
10	difficulties (EPs, PPs, CPs, HA) factors and RIIFA without items 7 and 23	2,210	974.551	128	<.001	.938	.925	.055	.056
11	EPs, CPs, HA reverse-worded removed, essential τ-equivalence	2,246	806.522	62	<.001	.919	.913	.073	.081

EPs = Emotional problems, PPs = Peer problems, CPs = Conduct problems, HA = Hyperactivity, PS = Prosocial behavior, obs = observations, df = degrees of freedom, RIIFA = Random Intercept Item Factor Analysis, item 7 = "Obedient," item 23 = "Adults"

Table 7. *Measurement invariance across genders and time using the model of difficulties factors and RIIFA without 7 and 23.*

Group	Groups	obs	Invariance	X^2	df	p	CFI	RMSEA	SRMR
2008: early adolescent	Girls vs. boys	3,503 / 3,375	configural	1726.356	257	<.001	.953	.041	.048
			weak	1662.168	274	<.001	.956	.038	.050
			strong	1799.345	287	<.001	.952	.039	.049
2013: early adolescent	Girls vs. boys	2,788 / 2,678	configural	1135.453	257	<.001	.961	.035	.045
			weak	1052.479	274	<.001	.965	.032	.047
			strong	1206.065	287	<.001	.959	.034	.046
early adolescent girls	T1 vs. T2	3,503 / 2,788	configural	1494.734	257	<.001	.957	.039	.046
			weak	1340.551	274	<.001	.963	.035	.046
			strong	1510.265	287	<.001	.958	.037	.046
early adolescent boys	T1 vs. T2	3503 / 2788	configural	1494.734	257	<.001	.957	.039	.046
			weak	1340.551	274	<.001	.963	.035	.046
			strong	1510.265	287	<.001	.958	.037	.046

obs = observations, df = degrees of freedom, RIIFA = Random Intercept Item Factor Analysis, item 7 = “Obedient,” item 23 = “Adults”

Table 8. *Girls' groups, reliability estimates.*

2008: early adolescent	**alpha α**	**ordinal alpha α**	**omega Ω**	**2008: mid-adolescent**	**alpha α**	**ordinal alpha α**	**omega Ω**
EPs	.697	.781	.715	EPs	.703	.781	.720
PPs	.534	.705	.548	PPs	.578	.735	.591
PPs without 23	.527	.720	.550	PPs without 23	.579	.749	.596
CPs	.519	.707	.546	CPs	.506	.658	.546
CPs without 7	.545	.741	.572	CPs without 7	.565	.719	.598
HA	.606	.706	.630	HA	.684	.760	.704
PS	.633	.750	.641	PS	.637	.748	.644
2013: early adolescent	**alpha α**	**ordinal alpha α**	**omega Ω**	**2013: mid-adolescent**	**alpha α**	**ordinal alpha α**	**omega Ω**
EPs	.674	.771	.690	EPs	.712	.793	.729
PPs	.524	.703	.545	PPs	.607	.764	.629
PPs without 23	.515	.712	.541	PPs without 23	.606	.777	.630
CPs	.457	.687	.497	CPs	.490	.694	.520
CPs without 7	.498	.735	.541	CPs without 7	.554	.750	.576
HA	.599	.705	.622	HA	.720	.794	.741
PS	.615	.749	.628	PS	.665	.786	.676

EPs = Emotional problems, PPs = Peer problems, CPs = Conduct problems, HA = Hyperactivity, PS = Prosocial behavior, item 7 = "Obedient," item 23 = "Adults"

Table 9. *Boys' groups, reliability estimates.*

2008: early adolescent	**alpha α**	**ordinal alpha α**	**omega Ω**	**2008: mid-adolescent**	**alpha α**	**ordinal alpha α**	**omega Ω**
EPs	.640	.763	.660	EPs	.644	.784	.675
PPs	.533	.694	.548	PPs	.589	.719	.598
PPs without 23	.550	.725	.568	PPs without 23	.575	.728	.600
CPs	.489	.661	.517	CPs	.511	.658	.538
CPs without 7	.527	.705	.550	CPs without 7	.582	.729	.592
HA	.560	.667	.590	HA	.633	.721	.657
PS	.653	.753	.655	PS	.659	.754	.663
2013: early adolescent	**alpha α**	**ordinal alpha α**	**omega Ω**	**2013: mid-adolescent**	**alpha α**	**ordinal alpha α**	**omega Ω**
EPs	.635	.768	.654	EPs	.701	**.819**	.719
PPs	.551	.726	.576	PPs	.642	.779	.666
PPs without 23	.568	.752	.596	PPs without 23	.619	.774	.644
CPs	.485	.697	.517	CPs	.553	.709	.596
CPs without 7	.524	.738	.551	CPs without 7	.661	**.806**	.674
HA	.593	.704	.615	HA	.663	.746	.696
PS	.662	.767	.667	PS	.660	.756	.662

EPs = Emotional problems, PPs = Peer problems, CPs = Conduct problems, HA = Hyperactivity, PS = Prosocial behavior, item 7 = "Obedient," item 23 = "Adults"

Table 10. *Spearman's rank correlation coefficients for sum score and factor score (4 factors RIIFA model without 7 and 23).*

SDQ Scale	2008 F early	2013 F early	2008 F mid	2013 F mid	2008 M early	2013 M early	2008 M mid	2013 M mid
EPs	.911	.906	.928	.934	.858	.845	.843	.824
PPs	.798	.770	.826	.814	.820	.782	.857	.817
PPs without 23	.886	.852	.912	.902	.908	.869	.929	.875
CPs	.753	.751	.805	.747	.768	.737	.798	.775
CPs without 7	.802	.804	.869	.821	.808	.772	.851	.822
HA	.947	.943	.974	.983	.927	.939	.971	.972

EPs = Emotional problems, PPs = Peer problems, CPs = Conduct problems, HA = Hyperactivity, PS = Prosocial behavior, obs = observations, df = degrees of freedom, RIIFA = Random Intercept Item Factor Analysis, item 7 = “Obedient,” item 23 = “Adults”