**PRE-PRINT: This is a pre-print version of the accepted manuscript and as such may contain errors. The paper is under copyright and the final version must be cited as:**
Chiorri, C., Hall, J., Casely-Hayford, J., & Malmberg, L.-E. (2016). Evaluating measurement invariance between parents using the Strengths and Difficulties Questionnaire (SDQ). *Assessment, 23*(1), 63–74. https://doi.org/10.1177/1073191114568301

# Evaluating Measurement Invariance between Parents using the Strengths and Difficulties Questionnaire (SDQ)

Carlo Chiorri[1*],

James Hall[2], Jeffrey Casely-Hayford[3] , Lars-Erik Malmberg[2]

[1]University of Genoa, Department of Educational Sciences, Genoa, Italy,

[2]University of Oxford, Department to Education, Oxford, UK,

[3]University of Bath, Department of Psychology, Bath, UK

Acknowledgments
The authors would like to thank the families who took part in the study and the research team. Data for this study were drawn from the Families, Children and Childcare Study, funded by the Tedworth Charitable Trust and the Glass-House Trust and led by Dr. Penelope Leach and Professor Jacqueline Barnes in London and by Professor Alan Stein, Professor Kathy Sylva and Dr. Lars-Erik Malmberg in Oxford (see www.familieschildrenchildcare.org). FCCC data are freely available to the research community. For information contact Professor Jacqueline Barnes, Department of Psychological Sciences, Birkbeck, University of London (jacqueline.barnes@bbk.ac.uk).

Word count (exc. abstract, reference list, tables and figures.): 5,899

*Correspondening Author:
Carlo Chiorri, PhD
University of Genova
Department of Educational Sciences
Corso A. Podestà, 2
16128 Genova, Italy
carlo.chiorri@unige.it

## Abstract

Parent ratings of their children's behavioral and emotional difficulties are commonly collected via the Strength and Difficulties Questionnaire (SDQ). For the first time, this study addressed the issue of inter-parent agreement using a measurement invariance approach. Data from 695 English couples (mothers and fathers) who had rated the behavior of their 4.25-year-old child were used. Given the inconsistency of previous results about the SDQ factor structure, alternative measurement models were tested. A 5-factor Exploratory Structural Equation Model (ESEM) allowing for non-zero cross-loadings fitted data best. Subsequent invariance analyses revealed that the SDQ factor structure is adequately invariant across parents, with inter-rater correlations ranging from .67 to .78. Fathers reported significantly higher levels of child conduct problems, hyperactivity, and emotional symptoms, and lower levels of prosocial behavior. This suggests that mothers and fathers each provide unique information across a range of their child's behavioral and emotional problems.

Word count of the abstract: 146


Key words: Measurement invariance, Exploratory Structural Equation Modeling, Multitrait Multisource Model; Strengths and Difficulty Questionnaire

# Evaluating Measurement Invariance between Raters using the Strength and Difficulties Questionnaire (SDQ)

## 1. Introduction

The Strengths and Difficulties Questionnaire (SDQ; Goodman, 1994) is a user-friendly instrument designed for the assessment of behavioral and emotional problems in children and adolescents aged 3-16 years. Its consists of 25 items equally divided across five scales, four of which probe difficulties: *emotional symptoms* (EMO), *conduct problems* (COND), *hyperactivity-inattention* (HYP), and *peer problems* (PEER); and one scale which probes strengths: *prosocial behavior* (PRO). The SDQ adopts a multi-informant approach, with versions designed for teachers and parents, and a self-report version for adolescents aged 11-16. Given its brief format, it is widely used as a screening tool in research, education, community, and clinical settings.

With translations into over 60 languages (Stone, Otten, Engels, Vermulst & Janssens, 2010) there has been much evaluative research on most of the psychometric properties of the SDQ. Factor structure, internal consistency**,** construct (convergent and discriminant) validity, and criterion (concurrent and predictive) validity have been extensively tested, but very few studies have investigated the inter-parent agreement of SDQ scores relating to their child. This is somewhat surprising, as in most studies a not-otherwise-specified 'parent' provided the data, without reporting whether it was the mother or the father (let alone those cases in which only one parent is available). Thus, inter-parent agreement is a crucial issue for all contexts in which the SDQ is used and where both parents are informants. If it can be shown that both parents provide the same information, then either can be confidently used as a single informant. On the other hand, if systematic differences exist between parents, then the use of both parental reports would

be advisable to enhance the sensitivity of scores in identifying children requiring clinical attention, as they add relevant unique variance (e.g., De Los Reyes & Kazdin, 2005).

One common and potentially problematic feature of studies on inter-parent agreement is that they usually rely on assumptions that might not be supported by empirical evidence. Analyses are usually carried out assuming that there is measurement invariance of the measures between mothers and fathers. To the best of our knowledge, no study has addressed this issue on the SDQ. This is important because in order to compare scores across parents, it must be shown that the latent dimensions that underlie the SDQ measure the same construct in the same way, and that the measurements themselves operate in the same way across parents: Otherwise, mean differences and other comparisons are likely to be invalid. The purpose of this study is thus to fill this gap in the literature.

*Inter-parent agreement on children's emotional and behavioral problems*

Past research on inter-parent agreement via reports of emotional and behavioral problems in children and adolescents has provided mixed results. For example, while a meta-analysis by Achenbach, McConaughy and Howell (1987) reported moderate, although significant, inter-parent agreement for both internalizing and externalizing problems, a similar study by Duhig, Renk, Epstein and Phares (2000) found a moderate correlation between mother and father ratings of internalizing problems, but higher inter-parent correlations for externalizing problems. Furthermore, inter-parent agreement varied with the age of the children: Achenbach et al. (1987) found a higher agreement for younger children while Duhig et al. (2000) found higher agreement for adolescents. In contrast, one consistent finding within the inter-parent agreement literature is that mothers consistently report more behavioral and emotional problems than do fathers,

although this may also depend on the measure employed and the age of the child (for reviews, see Davé, Nazareth, Senior & Sherr, 2008; Mellor, Wong, & Xu, 2011).

To the best our knowledge, only two studies have investigated the consistency and differences between mother and father ratings on the SDQ. Davé et al. (2008) collected data from 248 dyads composed of British biological mothers and fathers who were both residing with their own 4-to-6-year-old child. Cronbach's alphas were similar across all SDQ scales (PRO: .69 vs .70 for mothers and fathers, respectively; HYP: .72 vs .74; EMO: .54 vs 59; COND: .59 vs .57) but PEER (.58 vs .36), where the internal consistency of the maternal scale scores was significantly higher than that of the paternal scale scores. Spearman's correlations among raw scores of maternal and paternal scales were moderate to large (EMO: .39; COND: .51; HYP: .49; PEER: .41; PRO: .37) and fathers reported higher mean scores than mothers when rating HYP and COND (with small to moderate effect sizes).

Mellor et al. (2011) analyzed data from the parents of 700 primary school children (mean age: 8.7 years) in southwestern China. They focused on parental differences linked to children's gender and, like Davé et al. (2008), found similar Cronbach's alphas across parents and across all SDQ scales (EMO: Mothers .57 and .54 for girls and boys, respectively; Fathers: .57-.56; HYP: Mothers .66-.68; Fathers: .69-.68; PEER: Mothers .32-.29; Fathers: .29-.25; PRO: Mothers .66-.60; Fathers: .67-.61) other than for paternal ratings of boy's conduct problems (Mothers .42-.56; Fathers: .44-.40). Further analyses were then carried out after the scores on COND and HYP were combined into a single *externalizing problems* score (EXT), and excluding PEER due to its low reliability. Pearson correlations between mother and father scores were in the moderate to strong range: .61 and .59 for girls and boys, respectively (.40 and .46 for EMO and .46 and .38 for PRO). Further, mother and father reports of the behavioral difficulties of their daughters

agreed, whereas parents differed when rating the prosocial behavior of their sons (mothers reported significantly higher levels of prosocial behavior).

If we adopted the ratings suggested by Cicchetti and Sparrow (1981; 'poor' when lower than .40, 'fair' when ranging from .40 to .59, 'good' when ranging from.60 to .74 and 'excellent' when higher than .74) then the inter-parent agreement coefficients reported by Davé et al. (2008) and Mellor et al. (2011) would be considered 'poor' to 'fair'. Thus, the results of these two studies suggest only moderate agreement between parents on SDQ scores and this is consistent not only with previous research using other childhood measures of behavioral and emotional problems, but also with studies on other traits such as anxiety (Moreno, Silverman, Saavedra & Phares, 2008), psychopathology (De Los Reyes & Kazdin, 2005), Big Five personality (Tackett, 2011), and conflict and closeness in parent-child relationships (Driscoll & Pianta, 2011). These common inconsistencies are thought to reflect real differences in the way that a child is perceived by their parents and might stem from the tendency for mothers and fathers to play different parenting roles and to engage in different activities with a child. For example, mothers might have a higher participation in childrearing activities and spend more time with a child, particularly with infants and toddlers; moreover, a child might behave differently when alone with one of the parents, or the parents might promote different child behaviors (Davé et al., 2008; Driscoll & Pianta, 2011; Mascendaro, Herman, & Webster-Stratton, 2012; Mellor et al., 2011; Moreno et al., 2008; Tackett, 2011). However, inconsistent parent ratings seem to be more than just a result of differing parent roles. For example, it has also been shown that parental personal adjustment factors (stress, self-perceptions, substance abuse and marital discord) can contribute to parental attribution biases, especially in the assessment of their children’s externalizing problems (e.g., De Los Reyes, 2008; Liles et al., 2012).

*Testing measurement invariance on couple data*

Mother and father scores on the SDQ can be compared as long as it can be shown that the measures provided by the questionnaire are invariant across parents, i.e., the latent dimensions that underlie the SDQ measure the same construct in the same way, and that the measurements themselves operate in the same way across parents.

Testing the measurement invariance of parent observations of one child (i.e., within-couples) is different from testing the invariance of parents rating separate children. Instead of testing the same measurement model on two different groups as defined by parent's gender as a grouping variable, data have to be at the level of child, i.e., the unit of analysis is the child, and mothers and fathers are treated as different, but identifiable, raters of the same child. This means that, for each couple, both mother and father ratings are on the same line of data. An example of this modeling strategy has been provided by Burns et al. (2009), who evaluated measurement invariance between raters of the same child's Attention Deficit Hyperactivity Disorder (ADHD) and Oppositional Defiant Disorder (ODD). They proposed the application of Confirmatory Factor Analysis (CFA) within a *multiple indicator* (SDQ items) by *multitrait* (SDQ constructs) by *multisource* (mothers and fathers) model. This analysis tested if the measurement model for each SDQ factor (representing a scale score) remained invariant between mothers and fathers in all its parameters. More specifically, it tested whether: (1) there are differences in the reliability of scores between parents, (2) the extent of inter-parent agreement (i.e. same factor-different source correlations), (3) the degree of between-rater discriminant validity of the factors (i.e. if same factor-different source correlations are larger than the different factor-different source correlations), and (4) if parents perceived equal levels of each SDQ factor in their child (i.e. were

there mean differences?). However, a more recent study (Burns et al., 2013) questioned the use of CFA, as it requires each indicator to load on only one factor, thus assuming that secondary loadings are exactly zero (*Independent Cluster Model*, ICM, Church & Burke, 1994; or *perfect simple structure*, Sass & Schmitt, 2010). If one or more of the indicators have significant cross-loadings on a secondary factor, then the use of CFA might result in the failure to identify indicators with weak discriminant validity, i.e., with substantial loadings also on another factor or no substantial loadings on any factor. In the case of the SDQ, this would imply that a behavior that is considered an indicator of a specific problem can also be an indicator of another problem. In a CFA the more the cross-loadings depart from zero, the more the correlations among the SDQ factors will be inflated to account for non-zero cross-loadings restricted to zero, thus yielding: biased loadings, overestimated factor correlations, distorted structural relations, and lack of fit (Asparouhov & Muthén, 2009).

As pointed out by Booth and Hughes (2014), CFA can actually accommodate cross-loadings in models. If items are hypothesized to be complex and to measure multiple aspects of the construct under study, such paths can be specified *a priori*. Tests of their plausibility and consistency can then be carried out within a CFA framework. However, in some cases there might be no theoretical rationale that could inform the analyst when choosing the cross-loadings to be freed. In such a situation the analyst might revert to using modification indices for exploring and specifying a well-fitting measurement model instead of following the confirmatory route (Brown, 2001). This non-confirmatory positioning of the researcher is what leads to the use of Exploratory Factor Analysis (EFA). This technique appears preferable to CFA when searching for cross-loadings, since mis-specified loadings are easier to detect through rotation of the factor matrix than through the examination of modification indices in the case of CFA. Moreover, as

the process of freeing of parameters following modification indices is data-driven, the analyst is more susceptible to capitalization on chance characteristics of the data, thus undermining the generalization of results (e.g., MacCallum, Roznowski & Necowitz, 1992).

In response to the problem of potential cross-loadings in CFA, Asparouhov and Muthén (2009) advocated the use of Exploratory Structural Equation Modeling (ESEM). In these models a given number of factors can be specified grounding on *a priori* assumptions such that each item will have as many secondary loadings as there are factors. Thus, researchers can investigate the potential for cross-loadings so as to minimize bias. Note that ESEM is different from EFA as ESEM allows for an exploration of complex factor structures while allowing access to parameter estimates, standard errors, goodness-of-fit statistics, modeling flexibility (e.g., correlating error variances, obtaining factor scores corrected for measurement error, testing measurement invariance, etc.) – all features that are otherwise associated with CFA. Although less parsimonious than CFA models with the same number of factors, an ESEM model is a viable alternative to a CFA model when the CFA model is unable to fit the data. Relevant to the present study, Burns et al. (2013) proposed an application of a multiple indicator by multitrait by multisource model using ESEM to test the invariance across raters (mothers, fathers and teachers) of two measures of disruptive behavior in children and adolescents. Their findings indicated that ESEM can be a more appropriate procedure in those cases in which there might be indicators with strong loadings on a secondary factor that cannot be specified *a priori* on sound theoretical grounds.

*The present study*

We organized our study as follows: First, since previous research is inconclusive about the best factor structure for the SDQ (see Section 1 of Supporting Information for a comprehensive review of the measurement models to date), we tested the model fit of a series of theoretically plausible measurement models using CFA and ESEM for mothers and fathers separately. This was done to avoid the potential of testing invariance across an inadequately fitting measurement model. If we had not carried out this first stage we could not have known whether any possible lack of fit (when it came to invariance testing) was due to an actual lack of invariance or due to a poor choice of the measurement model. Second, we tested five forms of invariance of the best fitting model which allowed us to establish inter-parent consistencies and differences on SDQ scores.

## 2. Method

*Participants*

The sample for this study was UK-based and was drawn from the Families, Children and Child Care study (FCCC, www.familieschildrenchildcare.org; Sylva et al., 2007). Sampling centered on antenatal clinics and community post-natal clinics in Northern London and Oxfordshire, catering for a demographically diverse population. The recruited families came from a wide range of socio-economic backgrounds, and the attained sample was roughly comparable to the UK population at the time (see Malmberg et al., 2005).

Children in the study were followed up at 3, 10, 18, 36 and 51 months. Information was collected by face-to-face interviews with mothers (all data collection waves), and questionnaires to both parents (all data collection waves). After recruitment, the sample sizes of children were 1077,1050, 1036, and 1040, at 10, 18, 36 and 51 months respectively. For the present study we

used a subsample of 695 children who had both mother and father SDQ-ratings available when they were 4 years and 3 months (i.e., 51 months) old.

### *Measures*

Mothers and fathers completed the Strength and Difficulties Questionnaire (SDQ; Goodman, 1994). Each item was scored 0 = not true, 1 = somewhat true, and 2 = certainly true. Complete descriptive statistics are reported in Section 2 of the Supporting Information.

### *Analytic strategy*

#### Missing data procedure

The sample had negligible item-level missingness (0.5% missing data-points). As we dealt with ordered categorical indicators, we used the weighted least squares treatment of missing data implemented in Mplus (Asparouhov & Muthén, 2010), which is analogous to Full Information Maximum Likelihood (FIML). This method uses all of the observed data to produce parameter estimates that maximize the probability of the observed data having come from the population implied by those estimates.

#### SDQ Factor Structure

Following the literature, we specified CFA models with *one* (Behavioral Problems), *three* (Externalizing, Internalizing, and Prosocial Behaviors) and *five* (the original EMO, COND, HYP, PEER and PRO) constructs. Further, we also specified CFA models with *two* factors, as these would represent more parsimonious models with respect to three-factor models (i.e., Externalizing and Internalizing factors lumped together into a single Difficulties factor). In

multiple factor models, latent dimensions were specified to either correlate or not. We also considered the possibility of a method factor (Dickey & Blumberg, 2004; Palmieri & Smith, 2007; McCrory & Layte, 2012) that the author of the SDQ, Goodman (1994), identified as a "positive construal" factor, i.e., the general extent to which each parent tends to attribute positive qualities to the child, which might explain the covariance among items describing positive behaviors over and above that accounted for by trait factors (see Section 3 of the Supporting Information)

In the SDQ, positive behaviors are operationalized by the items of the PRO scale but also by two items in the PEER scale (item 11 "Has at least one good friend" and item 14 "Generally liked by other children"), two items in the HYP scale (item 21"Can stop and think things out before acting" and item 25 "Sees tasks through to the end, good attention span") and one item in the COND scale (item 7 "Generally obedient, usually does what adults request"). The specification of a method factor is a way to address the issue of a residual covariance among the items that is otherwise not accounted for by the latent dimensions that they are supposed to reflect. Previous results have shown that including a method factor in analyses of the SDQ's factor structure improves model fit (McCrory & Layte, 2012). That said, method factors can also wreck havoc in statistical models by causing problems such as non-convergence, improper solutions (i.e., parameter estimates out of range such as negative variance estimates or factor correlations greater than 1.0), and admissibility problems (due to empirical under-identification), especially when a method factor is correlated with the substantive factors (Marsh & Grayson, 1995). This is why we examined the factor structure of the SDQ both with, and without, method factors.

As an alternative to CFA, we also used 2- to 5-factor ESEM models to test the significance of cross-loadings. As explained in the introduction, ESEM models allowed for the testing of cross-loadings while providing more-accurate estimates of factor loadings, factor correlations and latent means. ESEM models with 2 to 5 factors (Geomin rotated) were also compared with CFA models with the same number of intercorrelated factors.

The Mplus software (Muthén & Muthén, 1998-2012) was used to test all models using the Weighted Least Squares Mean and Variance adjusted (WLSMV) estimator and Theta parameterization, which takes into account the non-normal categorical nature of item scores (as in Sanne, Torsheim, Heiervang, & Stormark, 2009). In order to determine whether a common measurement model could hold for mothers and fathers, we fitted all the CFA and ESEM models separately for mothers and fathers. The goodness-of-fit of the CFA and ESEM models was assessed using Bentler's Comparative Fit Index (CFI; Bentler, 1990), the Tucker Lewis Index (TLI; Tucker & Lewis, 1973), the Root-Mean-Square Error of Approximation (RMSEA; Steiger & Lind, 1980), and the Weighted Root-Mean-square Residual (WRMR; Yu, 2002). Although we also report chi-square values, it must be noted that they cannot be straightforwardly evaluated when using WLSMV since degrees of freedom are estimated "using a diagonal weight matrix with standard errors and mean- and variance adjusted chi-square test statistic that use a full weight matrix" (Muthén & Muthén, 1998-2012, p. 603). Following Yu (2002), we used the following cutoff values as indicative of good fit: CFI $\geq$ .96, TLI $\geq$ .95, RMSEA $\leq$ .05, and WRMR $\leq$ 1.00.

SDQ Measurement invariance models

After finding a common measurement model for mothers and fathers, we then used this model in a test of the invariance of parental ratings. Traditionally (e.g., Meredith, 1993), the sequence of invariance testing begins with a 'configural invariance' model in which all parameters are freely estimated, such that the only similarity of the overall pattern of parameters is evaluated. Technically, this model is not an invariance model in that it does not require any estimated parameters to be the same. Nevertheless, it is necessary since it provides (i) a test of the ability of the *a priori* model to fit the data in each child's Parental Rater without invariance constraints and (ii) a baseline for comparing other models that impose equality constraints on the parameter estimates across raters of the same child.

Tests of configural invariance models were followed by tests of *weak invariance*, which constrain factor loadings to be invariant over raters. If identical items have statistically equivalent loadings, then the scores of identical items show the same amount of increase between raters for the same amount of increase on the latent factor (i.e., equality of scaling units). However, changes in the means of the latent factors can only be interpreted as changes in the latent constructs if the indicator metric is invariant across raters (*strong invariance*). In other words, raters with the same level of the latent construct will have the same expected score of the measured indicators. In the case of ordinal indicators, item scores are assumed to reflect the amount of an underlying continuous and normally distributed variable (Muthén & Asparouhov, 2002). The *strong invariance* model tests whether the thresholds (or distribution cut points, i.e., *z*-values corresponding to the cumulative probability for each response category except the highest) are equal across raters. In other words, threshold invariance is satisfied when the cut points on the underlying normal distribution are equal across raters for each indicator. If the

*strong invariance* model shows a fit similar to the weak invariance model, it means that raters with the same level of the latent construct will have the same expected score on the measured indicators. This also implies that any observed score differences between raters on identical items is not due to rater bias but rather due to true differences on the latent construct mean. However, although *strong invariance* allows for testing differences in *latent* means, it is insufficient for testing difference in *manifest* (raters') means, which also require items' residual variances to be invariant. The presence of differences in reliability (as represented in the item residual variances) across raters could in fact distort mean differences on the observed scores. In the case of ordinal indicators, this means that the estimates of the residual variances of the continuous and normally distributed variables underlying item scores are constrained to be equal.

As mentioned earlier, when dealing with paired samples and non-normal categorical indicators some issues must be taken into account. Since the unit of analysis was the child, and mothers and fathers were treated as different and known (i.e., non-exchangeable) raters of the same child, we specified the equivalent of a single-group correlated-factor model in which the items of the scale were considered twice, as indicators of mother and father perceptions of their child's behavior. Each child therefore had 50 symptom ratings, 25 rated by the mother and 25 by the father, with correlations between corresponding factors being a measure of inter-parent agreement. This also implies that the systematic residual variance (uniqueness) in each pair of identical items between parents is expected to covary because of the identical nature of each item pair. For example, the residual variance in the item "Generally liked by other children" for mothers should covary with the item "Generally liked by other children" for fathers. Hence, we allowed correlated residual variances between like items (Figure 1).

[Figure 1]

We set the identification constraints of invariance models to the values suggested by Muthén and Muthén (1998-2012, p. 486). When using the WLSMV estimator and Theta parameterization for observed categorical indicators, residual variances of the latent response variables underlying the observed categorical indicators are part of the default model. The residual variances of both groups cannot be simultaneously estimated, but the first group has residual variances fixed at one for all observed categorical indicators and in the other group the residual variances are free to be estimated with starting values of one (Muthén & Asparouhov, 2002). The least restrictive model ('configural invariance': *Model 1*) is thus a model in which item thresholds and factor loadings are free across groups; residual variances are fixed at one in all groups; factor means are fixed at zero in all groups. Equality constraints were then added to model parameters to test different degrees of invariance. In *Model 2*, factor loadings were constrained to equality (but note that residual variances are still invariant due to identification issues), while in *Model 3* equality of thresholds was added. Note that in this latter model we also freed residual variances and latent means in one group. In *Model 4*, the residual variances of the latent observed variables in both groups were fixed at 1. Latent means invariance was finally tested (by fixing them to zero in both groups) in *Model 5*.

When comparing statistical models we considered more parsimonious models to be supported as preferable if there was a difference between the fit of models of less than .01 on the CFI (Chen, 2007) or a difference in RMSEA of less than .015 (Chen, 2007). Since Marsh (2007) noted that some indices (e.g., TLI and RMSEA) incorporate a penalty for lack of parsimony so that the more parsimonious model fits data better than a less parsimonious model (i.e., the gain in

parsimony is greater than the loss in fit), we also considered the more parsimonious model to be supported by a TLI or RMSEA which was as good as, or better than, a more complex model. Note that the Bayesian Information Criterion (BIC) contains a more appropriate parsimony penalty for comparing the CFA and ESEM models, but it cannot be computed when using WLSMV estimation, since it needs the log-likelihood value, which can be obtained only through maximum likelihood estimation.

## 3. Results

### *SDQ Factor Structure*

Results of the CFAs and ESEMs for mothers and fathers are reported in Table 1.

[Table 1]

Among the CFA models, the model specifying five correlated factors and a positive construal method factor (Model 14 in Table 1) had the best fit for both mothers and fathers, but with CFI and TLI substantially (i.e., > .01) lower than optimal values for the father model. Modification indices suggested that the lack of fit could be ascribed to significant cross-loadings on substantive factors and, even more problematically for the interpretation of the model, to significant loadings on the method factor of non-positive items.

As pointed out in the introduction, in a CFA framework the post-hoc specification of cross-loadings only grounding on modification indices might be problematic in terms of the generalization of results. Hence, we opted for ESEM. The 5-factor model had the best fit among the ESEM models. Hence, we concluded that the 5-factor ESEM model should be used in

subsequent invariance tests. Note however that an adequate fit of an ESEM model does not necessarily mean what Sass and Schmitt (2010) call an "*approximate simple structure"*, i.e., that each item has a substantial loading on one factor and negligible loadings (i.e., < |.30|) on the others (cross-loadings). Table 5 in Section 4 of the Supporting Information shows this is not the case, and the presence of substantial cross-loadings explains the higher fit of the ESEM model with respect to the CFA 5-correlated-factor model. One scale which appears problematic is HYP as (for both parents) two items (2, 10) had substantial cross-loadings on COND and two other items (21, 25) on PRO. Moreover, two items of PEER (11, 14) also loaded on PRO and one item of COND (5) also loaded on EMO.

The reliability of latent scores in ESEM models was computed as the composite reliability index (Raykov, 1997). Values for COND were mothers=.77 and fathers=.82 (difference *z*-test: $p$=.001); for EMO mothers=.82 and fathers=.73 ($p$<.001); for HYP mothers=.79 and fathers=.76 ($p$=.053); for PEER mothers=.66 and fathers=.75 ($p$<.001); for PRO mothers=.84 and fathers=.84 ($p$=.627).

*Invariance of mother and father ratings*

In testing the invariance of the 5-factor ESEM model, cross-loadings were allowed *only* within each source. In other words, an item rated by the mother was allowed to have non-zero cross-loadings on all other mother factors, but not on father factors, and vice versa. Correlated residual variances were specified between identical mother- and father-rated items *a priori* (see Figure 1). The fit of the invariance models did not substantially decrease when imposing equality constraints ($\Delta$CFI, $\Delta$TLI > .01, $\Delta$RMSEA > .015; see Table 2), suggesting that the 5-factor ESEM model is reasonably invariant across mothers and fathers.

[Table 2]

However, the inspection of parameters representing factor means differences in Model 4 revealed that fathers tended to endorse higher scores in COND (standardized coefficient = .13, $p$ = .009), HYP (.14, $p$ = .001), EMO (.16, $p$ = .002) and that mothers tended to endorse higher scores in PRO (.10, $p$ = .046). No difference was observed in PEER (.02, $p$ = .743) .

It must be noted that Mplus does not allow constraints to be placed on the variance of ESEM factors in single-group analyses. As our model was equivalent to a single-group ESEM model with 50 items and 10 factors (see Figure 1 and the description of the model above), this meant that the invariance of factor variances and inter-correlations could not be tested by comparing a model with equality constraints on factor correlations against Model 4.

*Inter-parent agreement on SDQ scores*

Table 3 reports the correlations from tests of the strict invariance model (Model 4 in Table 2). The *same* factor-different source correlations (i.e., agreement between parents) were larger (range .67 to .78, median .70) than the *different* factor-different source correlations (range -.20 to .28, median absolute value .11), suggesting adequate discriminant validity. Grounding on the guidelines of Cicchetti and Sparrow (1981), the same factor-different source correlations suggested a 'good' agreement for all scales but PEER, where the agreement was 'excellent'.

[Table 3]

## 4. Discussion

The aim of this study was to investigate agreement and differences between mother and father reports of their child's behavioral and emotional problems as assessed by the SDQ. To address this aim, we conducted a multiple indicator (problematic behaviors) by multitrait (five latent dimensions) by multisource (mothers and fathers of the same child) invariance analysis. Demonstrating measurement invariance between mother and father ratings of the same child is necessary for studies which use parental ratings on the SDQ to draw valid conclusions about inter-rater reliability and mean score differences. In turn, this would allow researchers to draw conclusions about whether both parents provide the same information, and thus whether either can be confidently used as a single informant. Alternatively, a lack of inter-rater agreement and/or substantial mean score differences would suggest that mothers and fathers would be providing different perspectives and thus potentially relevant and unique information about their child's behavioral and emotional problems.

We began with an investigation into the SDQ factor structure, an issue which had not been conclusively addressed by previous research (see Section 1 of the Supporting Information). We found that in both parents an ESEM model allowing for non-zero cross-loadings fitted substantially better than both a 5-correlated-factor CFA model and a 5-correlated-factor CFA model with an additional positive construal method factor. This means that there was only weak support for the model that is commonly used by researchers to compute scale scores (a 5-factor ICM-CFA model). The problem is that some items, especially in the HYP scale, appeared to be indicators of more than one construct. This result has two important implications: (1) Additional work on the content and the wording of the SDQ items might improve the validity of the questionnaire; (2) If cross-loadings consistent with an ESEM approach are required to fit the

data, then a simple unweighted average of the multiple indicators (based on ICM) is unlikely to provide an optimal representation of the latent construct (Marsh et al., 2009). Hence, the results of this study suggest if the SDQ is part of a Structural Equation Model (e.g., a latent growth model to investigate systematic change in children's behavior), then the analyses will be more appropriately carried out via use of an ESEM approach than a traditional ICM-CFA approach. However, it must be noted that this does not mean that the ESEM approach should *always* replace the corresponding CFA approach. When a more parsimonious CFA model fits the data as well as the ESEM model does, then the CFA should be used. And even when the CFA does not adequately fit the data, if items are hypothesized to be complex and to measure multiple aspects of the construct under study, cross-loadings can be specified *a priori* and their plausibility and consistency tested while still using a CFA framework (Booth & Hughes, 2014). However, when there are no theoretical grounds to support the specification of cross-loadings, and thus when researchers are obliged to rely on post-hoc modification indices, ESEM models might provide a viable alternative to CFA.

With the preferable factor structure of the SDQ established, the central aim of this paper could then be addressed. The results of the parent-rating invariance analysis showed that a 5-factor ESEM model of the SDQ was reasonably invariant across parents. Although some differences in scale internal consistency were found when considering scores separately for mothers and fathers (see Table 2), the negligible loss of fit for the invariance model that constrained to equality factor loadings, thresholds, and residual variances suggested negligible differences in the reliability of mother and father scores.

The demonstration of the invariance of item loadings and thresholds then allowed a valid evaluation of inter-parent agreement and of the invariance of the factor means between sources.

Estimates of the inter-parent agreement were all 'good' and one 'excellent'. Further, the level of agreement did not differ between internalizing and externalizing behaviors. This finding is inconsistent not only with previous research on the SDQ (Davé et al., 2008; Mellor et al., 2011), but also with research on other psychological measures such as the Big Five personality types (higher agreement on more easily observable traits such as Conscientiousness and Openness to Experience than on an internal, less observable trait such as Neuroticism; Tackett, 2011) and parent-child relationships (higher agreement for conflict than for closeness, Driscoll & Pianta, 2011).

In contrast, a finding that *was* consistent with literature was that fathers' scores were higher than mothers' for COND, HYP and EMO and lower for PRO. While future research should shed light on the reasons for these results, they suggest that mothers and fathers provide different and unique perspectives in reporting on their child's behavioral and emotional problems, and thus, whenever possible, they should both be collected. Note that this does not necessarily mean that scores should be averaged. As suggested by Tackett (2011), when utilizing mean-level ratings to predict later behavior or to guide assessment and treatment, the presence of discrepancies on child's personality ratings can create confusion, as it might be an indicator of conflict in the family system and might point to other sources of clinically relevant information that could be useful in case conceptualization and treatment planning.

## Limitations and future research

Some limitations of this study must be acknowledged. First, although the families in the FCCC-study were fairly representative of the areas they were sampled, (Malmberg et al., 2005) the subsample here (both mother *and* father ratings of the same child) excluded single-parent

families *a priori*. Stable couples were likely to be more advantaged than single-parent or restructured families. As child problem behavior is more prevalent among disadvantaged families, so the range of the scores here reaches clinical levels in no more than 3% of cases (see also Stein et al., 2012 and Supporting Information, Section 2, Table 4).

Substantively, although our findings suggest that mother and father ratings do not seem to be fully interchangeable, the results at this point might be considered specific to: (1) cultural context (limited to UK), (2) sampling of parent dyads (no data on non-stable couples was considered), (3) age range (limited to 51 months). Given the worldwide availability of the SDQ, it would be useful to repeat the analyses reported in this paper in different cultural contexts, with non-stable couples and with other age ranges.

Methodologically, the exploration and invariance aspects of the analyses have been conducted on the same sample, but testing invariance on an independent sample would have provided stronger evidence. As reported by Burns et al. (2013), some limitations of the ESEM multiple indicators by multitrait by multisource model must also be pointed out. This model cannot separate variability in the individual behavior ratings into latent source and latent trait effects. In other words, it cannot determine how much of the variance in the behavior ratings for mothers and fathers is trait variance, source variance, and residual. If answering research questions requires the specification of latent source and trait factors in order to relate these factors to predictors and outcomes, then a 'multiple indicator by correlated trait by correlated method minus one model' would allow for a better examination of trait and source effects (see Eid, Lischetzke & Nussbeck, 2006). Moreover, Dumenci, Achenbach and Windle (2011) suggested model to measure context-specific and cross-contextual effects in multiple source rating scales.

Future studies could also include other ways in which invariance could be assessed, for example if mothers and fathers rate the behaviors of boys and girls differently. Several studies have found an interaction between the gender of parent and the gender of the rated child, whereby mothers report greater problems for sons than do fathers, and fathers report more problems for daughters than do mothers (Stanger & Lewis, 1993; Duhig et al., 2000). Jensen et al. (1988) reported that mothers and fathers differed significantly in their ratings of their sons' behavioral problems, but not their daughters', with mothers reporting more problems for their sons., but other studies (e.g., Achenbach, Howell, Quay & Connors, 1991; Stanger & Lewis, 1993) have found no parent-gender by child-gender interaction in ratings of behavioral problems. For the SDQ, Davé et al. (2008) found that fathers were significantly more likely to report conduct problems, compared to mothers, among their daughters, while Mellor et al. (2011) reported that mothers endorsed significantly higher scores than fathers for prosocial behaviors for their sons. In principle it is possible to specify an ESEM-within-CFA model that partitions latent mean differences into tests of rater, child gender, and interaction effects (e.g., Marsh, Nagengast & Morin, 2012). In this study however, while all scoring categories were endorsed at least once for each item in the total sample used for the analyses (see Section 2, Table 2 of the Supporting Information), we found that in some items the highest scoring category was never endorsed in either the boy or girl subgroups. This would not have allowed us to test such models without resorting to data transformations. For example, by collapsing the two highest scoring categories.

## Conclusions

In this paper we have shown the usefulness of ESEM in investigating interparent agreement on the SDQ. Results led to the conclusion that although mothers and fathers report on the same problems, they do not necessarily report the same level of problems (fathers had a tendency to report more difficulties and fewer strengths). This suggests that when possible, ratings from both parents should be collected as they provide unique information on their child's behavioral and emotional problems.

**References**

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child-adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101*, 213-232. doi:10.1037-0033-2909.101.2.213

Achenbach, T. M., Howell, C. T., Quay, H. C., & Connors, C. K.(1991). National survey of problems and competencies among four- to sixteen-year olds: Parents' reports for normative and clinical samples. *Monographs of the Society for Research in Child Development, 56* ,7-11. Available at http://www.jstor.org/stable/1166156 [November 23rd, 2014].

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

Asparouhov, T., & Muthén, B. (2010). *Weighted least squares estimation with missing data*. Available online at: http://www.statmodel.com/download/GstrucMissingRevision.pdf [November 23rd, 2014].

Bentler, P.M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-246. doi: 10.1037/0033-2909.107.2.238

Booth, T., & Hughes, D. J. (2014). Exploratory Structural Equation Modeling of Personality Data. *Assessment, 21*(3), 260-271. doi: 10.1177/1073191114528029

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research, 36,* 111-150. doi: 10.1207/S15327906MBR3601_05

Burns, G. L., Desmul, C., Walsh, J. A., Silpakit, C., & Ussahawanitchakit, P. (2009). A multitrait (ADHD–IN, ADHD–HI, ODD toward adults, academic and social competence) by multisource (mothers and fathers) evaluation of the invariance and convergent/discriminant validity of the Child and Adolescent Disruptive Behavior Inventory with Thai adolescents. *Psychological Assessment, 21*(4), 635-641. doi: 10.1037/a0016953

Burns, G. L., Walsh, J. A., Severa, M., Lorenzo-Seva, U., Cardo, E., & Rodríguez-Fornells, A. (2013). Construct validity of ADHD/ODD rating scales: Recommendations for the evaluation of forthcoming DSM-V ADHD/ODD scales. *Journal of Abnormal Child Psychology, 41,* 15-26. doi: 10.1007/s10802-012-9660-5

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

Church, A. T., Burke, P. J. (1994). Exploratory and confirmatory tests of the Big Five and Tellegen's three-and four- dimensional models. *Journal of Personality and Social Psychology, 66,* 93–114. doi: 10.1037/0022-3514.66.1.93

Cicchetti, D. V. & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency, 86,* 127-137.

Davé, S., Nazareth, I., Senior, R., & Sherr, L. (2008). A Comparison of father and mother report of child behaviour on the Strengths and Difficulties Questionnaire. *Child Psychiatry & Human Development, 39,* 399-413. doi: 10.1007/s10578-008-0097-6

De Los Reyes, A. (2008). Whose depression relates to discrepancies? Testing relations between informant characteristics and informant discrepancies from both informants' perspectives. *Psychological Assessment, 20,* 139-149. doi: 10.1037/1040-3590.20.2.139

De Los Reyes, A., & Kazdin, A. E. (2005). Informant discrepancies in the assessment of childhood psychopathology: A critical review, theoretical framework, and recommendations for further study. *Psychological Bulletin, 131,* 483-509. doi:10.1037/0033-2909.131.4.483

Dickey, W., C., & Blumberg, S., J. (2004). Revisiting the factor structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of the American Academy of Child and Adolescent Psychiatry, 43,* 1159-1167. doi: 10.1097/01.chi.0000132808.36708.a9

Driscoll, K., & Pianta, R. C. (2011). Mothers' and fathers' perceptions of conflict and closeness in parent-child relationships during early childhood. *Journal of Early Childhood and Infant Psychology, 7,* 1-20.

Duhig, A.M., Renk, K., Epstein, M.K., & Phares, V. (2000). Inter-parental agreement on internalizing, externalizing, and total behavior problems: A meta-analysis. *Clinical Psychology: Science and Practice, 7,* 435- 453. doi: 10.1093/clipsy.7.4.435

Dumenci, L., Achenbach, T. M., & Windle, M. (2011). Measuring context-specific and cross-contextual components of hierarchical constructs. *Journal of Psychopathology and Behavioral Assessment, 33,* 3-10. doi: 10.1007/s10862-010-9187-4

Eid, M., Lischetzke, T., & Nussbeck, F. W. (2006). Structural equation modeling for multitrait-multimethod data. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 283-299). Washington, DC: American Psychological Association.

Goodman, R. (1994). A modified version of the Rutter parent questionnaire including items on children's strengths: A research note. *Journal of Child Psychology and Psychiatry, 35*, 1483–1494. doi: 10.1111/j.1469-7610.1994.tb01289.x

Jensen, P.S., Traylor, J., Xenakis, S.N., & Davis, H. (1988). Child psychopathology rating scales and interrater agreement: I. Parents' gender and psychiatric symptoms. *Journal of the American Academy of Child & Adolescent Psychiatry. 27,* 442-450. doi: 10.1097/00004583-198807000-00012

Liles, B. D., Newman, E., Lagasse, L. L., Derauf, C., Shah, R., Smith, L. M., … Lester, B. M. (2012). Perceived child behavior problems, parenting stress, and maternal depressive symptoms among prenatal methamphetamine users. *Child Psychiatry and Human Development, 43*, 943–957. doi:10.1007/s10578-012-0305-2

MacCallum, R.C., Roznowski, M., & Necowitz, L.B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin, 111*(3), 490-504. doi: 10.1037/0033-2909.111.3.490

Malmberg, L.-E., Davies, B., Walker, J., Barnes, J., Sylva, K., Stein, A., & Leach, P. (2005). *The Families, Children and Child Care (FCCC) study in relation to area characteristics: Recruitment and sample description*. Available online at: http://www.familieschildrenchildcare.org/fccc_static_PDFs/fccc_sample_recruit.pdf [November 23rd, 2014].

Marsh, H. W. (2007). Application of confirmatory factor analysis and structural equation modeling in sport/exercise psychology. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (3rd ed., pp. 774 –798). Hoboken, NJ: Wiley.

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J. S., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFAand EFA: Application to students' evaluations of university teaching. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(3), 439-476. doi:10.1080/10705510903008220

Marsh, H. W., Nagengast, B., Morin, A. J. S. (2012). 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., & Grayson, D. (1995). Latent variable models of multitrat-multimethod data. In Hoyle, R.H (ed.) *Structural equation modelling: Concepts, issues and applications*. 177-198. Thousand Oaks, Ca, US: Sage Publications.

Mascendaro, P. M., Herman, K. C., & Webster-Stratton, C. (2012). Parent discrepancies in ratings of young children's co-occurring internalizing symptoms. *School Psychology Quarterly*, *27*, 134-143. doi:10.1037/a0029320

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., Wong, J., & Xu, X. (2011). Inter-parent agreement on the Strengths and Difficulties Questionnaire: A Chinese study. *Journal of Clinical Child & Adolescent Psychology*, *40*, 890-896. doi: 10.1080/15374416.2011.614580

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543. doi: 10.1007/BF02294825

Moreno, J., Silverman, W. K., Saavedra, L. M., & Phares V. (2008). Fathers' ratings in the assessment of their child's anxiety symptoms: a comparison to mothers' ratings and their

associations with paternal symptomatology. *Journal of Family Psychology, 22*, 915-919. doi: 10.1037/a0014097

Muthén, B. O., & Asparouhov, T. (2002). *Latent variable analysis with categorical outcomes: multiple-group and growth modeling in Mplus*. Available online at: http://www.statmodel.com/download/webnotes/CatMGLong.pdf . Mplus Web Notes: No.4. [November 23rd, 2014].

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide. 7thEdn*. Los Angeles, CA: Muthén and Muthén.

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

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*, 329-353. doi:10.1207/s15327906mbr3204_2

Sanne, B., Torsheim, T., Heiervang, E., & Stormark, K. M. (2009). The Strengths and Difficulties Questionnaire in the Bergen Child Study: A conceptually and methodically motivated structural analysis. *Psychological Assessment, 21*, 352-364. doi: 10.1037/a0016317

Sass, D. A., & Schmitt, T. A. (2010). A comparative investigation of rotation criteria within exploratory factor analysis. *Multivariate Behavioral Research, 45*, 73–103. doi: 10.1080/00273170903504810

Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. *Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA*.

Stein, A., Malmberg, L-E., Sylva, K., Leach, P., Barnes, J., & FCCC (2012). The influence of different forms of early childcare on children's emotional and behavioural development at school entry. *Child: Care, Health & Development, 39,* 676-687. doi: 10.1111/j.1365-2214.2012.01421.x

Stone, L. L., Otten, R., Engels, R. C., Vermulst, Ad A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: A Review. *Clinical Child and Family Psychology Review*, *13*, 254-274. doi: 10.1007/s10567-010-0071-2

Stanger, C., & Lewis, M. (1993). Agreement among parents, teachers, and children on internalizing and externalizing behavior problems. *Journal of Clinical Child Psychology. 22,* 107-116. doi: 10.1207/s15374424jccp2201_11

Sylva, K., Stein, A., Leach, P., Barnes, J., Malmberg, L-E. & FCCC team. (2007). Family and child factors related to the use of infant care: An English study. *Early Childhood Research Quarterly*. *22*, 118-136. doi: 10.1016/j.ecresq.2006.11.003

Tackett, J. L. (2011). Parent Informants for Child Personality: Agreement, Discrepancies, and Clinical Utility. *Journal of Personality Assessment, 93*(6), 539-544. doi:10.1080/00223891.2011.608763

Tucker, L. R. & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1-10. doi: 10.1007/BF02291170

Yu, C. Y. (2002). *Evaluating cutoff criteria of model fit indices for latent variable models with binary and continuous outcomes.* Doctoral dissertation, University of California, Los Angeles. Available at: http://www.statmodel.com/download/Yudissertation.pdf [November 23rd, 2014].

*Table 1 Goodness of fit of Strengths and Difficulties Questionnaire Confirmatory Factor Analysis (CFA) and Exploratory Structural Equation Modeling (ESEM) measurement models. ESEM models are bolded for ease of interpretation (n=695)*

| Model | Rater | $\chi^2$ | df | CFI | TLI | RMSEA | WRMR |
|---|---|---|---|---|---|---|---|
| CFA model 1: 1 factor | Mother | 1581.792 | 275 | .672 | .642 | .083 | 2.274 |
| | Father | 1460.509 | 275 | .665 | .635 | .079 | 2.149 |
| CFA model 2: 1 factor + 1 Method factor | Mother | 1035.771 | 265 | .806 | .781 | .065 | 1.783 |
| | Father | 1028.707 | 265 | .784 | .756 | .064 | 1.752 |
| CFA model 3: 2 factors uncorrelated | Mother | 1654.304 | 275 | .654 | .622 | .085 | 2.480 |
| | Father | 1821.729 | 275 | .563 | .524 | .090 | 2.586 |
| CFA model 4: 2 factors uncorrelated + Method factor | Mother | 1112.929 | 265 | .787 | .759 | .068 | 1.958 |
| | Father | 1221.174 | 265 | .730 | .695 | .072 | 2.043 |
| CFA model 5: 2 factors correlated | Mother | 1286.025 | 274 | .746 | .722 | .073 | 2.041 |
| | Father | 1269.414 | 274 | .719 | .692 | .072 | 1.995 |
| **ESEM model: 2 factors** | **Mother** | **1081.769** | **251** | **.791** | **.751** | **.069** | **1.728** |
| | **Father** | **1019.188** | **251** | **.783** | **.741** | **.066** | **1.654** |
| CFA model 6: 2 factors correlated + Method factor | Mother | 1033.795 | 264 | .807 | .780 | .065 | 1.782 |
| | Father | 1027.600 | 264 | .785 | .755 | .065 | 1.752 |
| CFA model 7: 3 factors uncorrelated | Mother | 1348.480 | 275 | .730 | .706 | .075 | 2.370 |
| | Father | 1609.014 | 275 | .624 | .589 | .084 | 2.603 |
| CFA model 8: 3 factors uncorrelated + Method factor | Mother | 835.166 | 265 | .857 | .838 | .056 | 1.777 |
| | Father | 1070.989 | 265 | .773 | .742 | .066 | 2.042 |
| CFA model 9: 3 factors correlated | Mother | 857.099 | 272 | .853 | .838 | .056 | 1.634 |
| | Father | 939.705 | 272 | .812 | .792 | .059 | 1.695 |
| **ESEM model: 3 factors** | **Mother** | **513.351** | **228** | **.928** | **.906** | **.042** | **1.023** |
| | **Father** | **552.752** | **228** | **.908** | **.879** | **.045** | **1.086** |

*(continues)*

*Table 1 Goodness of fit of Strengths and Difficulties Questionnaire measurement models (ctd.)*

| Model | Rater | $\chi^2$ | df | CFI | TLI | RMSEA | WRMR |
|---|---|---|---|---|---|---|---|
| CFA model 10: 3 factors correlated + Method factor | Mother | 583.019 | 262 | .919 | .908 | .042 | 1.290 |
| | Father | 688.836 | 262 | .880 | .862 | .048 | 1.406 |
| **ESEM model: 4 factors** | **Mother** | **387.142** | **206** | **.955** | **.934** | **.036** | **0.850** |
| | **Father** | **376.299** | **206** | **.952** | **.930** | **.034** | **0.841** |
| CFA model 11: 5 factors uncorrelated | Mother | 1698.375 | 275 | .642 | .610 | .086 | 2.810 |
| | Father | 1883.620 | 275 | .546 | .505 | .092 | 2.955 |
| CFA model 12: 5 factors uncorrelated + Method factor | Mother | 1175.578 | 265 | .771 | .741 | .070 | 2.259 |
| | Father | 1357.412 | 265 | .692 | .651 | .077 | 2.431 |
| CFA model 13: 5 factors correlated | Mother | 621.234 | 265 | .911 | .899 | .044 | 1.327 |
| | Father | 695.485 | 265 | .879 | .862 | .048 | 1.405 |
| **ESEM model: 5 factors** | **Mother** | **266.020** | **185** | **.980** | **.967** | **.025** | **0.655** |
| | **Father** | **286.481** | **185** | **.971** | **.954** | **.028** | **0.695** |
| CFA model 14: 5 factors correlated + Method factor | Mother | 401.899 | 255 | .963 | .957 | .029 | 1.005 |
| | Father | 466.906 | 255 | .940 | .930 | .035 | 1.100 |
| CFA model 15: 5 factors +1 Higher Order Factor uncorrelated with prosocial | Mother | 1206.227 | 271 | .765 | .740 | .070 | 2.090 |
| | Father | 1426.813 | 271 | .674 | .639 | .078 | 2.279 |
| CFA model 16: 5 factors+ 1 Higher Order Factor correlated with prosocial | Mother | 708.834 | 270 | .890 | .878 | .048 | 1.462 |
| | Father | 775.440 | 270 | .857 | .842 | .052 | 1.520 |

Note: df=Degrees of Freedom; CFI=Comparative Fit Index; TLI=Tucker–Lewis Index; RMSEA= Root-Mean-Square Error of Approximation; WRMR= Weighted Root-Mean-square Residual.

*Table 2 Goodness of fit of measurement invariance Exploratory Structural Equation Models fot the Strengths and Difficulties Questionnaire (n = 695)*

| Invariance Model | FL | TH | RV | M | $\chi^2$ | df | CFI | TLI | RMSEA | WRMR |
|---|---|---|---|---|---|---|---|---|---|---|
| Model 1 | | | X | X | 1112.411 | 945 | .980 | .974 | .016 | 0.737 |
| Model 2 | X | | X | X | 1177.236 | 1045 | .984 | .981 | .013 | 0.817 |
| Model 3 | X | X | | | 1224.117 | 1065 | .981 | .978 | .015 | 0.830 |
| Model 4 | X | X | X | | 1250.411 | 1090 | .981 | .978 | .015 | 0.851 |
| Model 5 | X | X | X | X | 1285.049 | 1095 | .977 | .974 | .016 | 0.872 |

Note: FL=factor loadings; TH=thresholds; RV=residual variances; M=factor means; Xs indicate that the parameter is invariant across raters; df=Degrees of Freedom; CFI=Comparative Fit Index; TLI=Tucker–Lewis Index; RMSEA= Root-Mean-Square Error of Approximation; WRMR= Weighted Root-Mean-square Residual.

*Table 3 Multitrait by multisource factor correlation matrix from Model 4 in Table 3 for the Strengths and Difficulties Questionnaire (n = 695)*

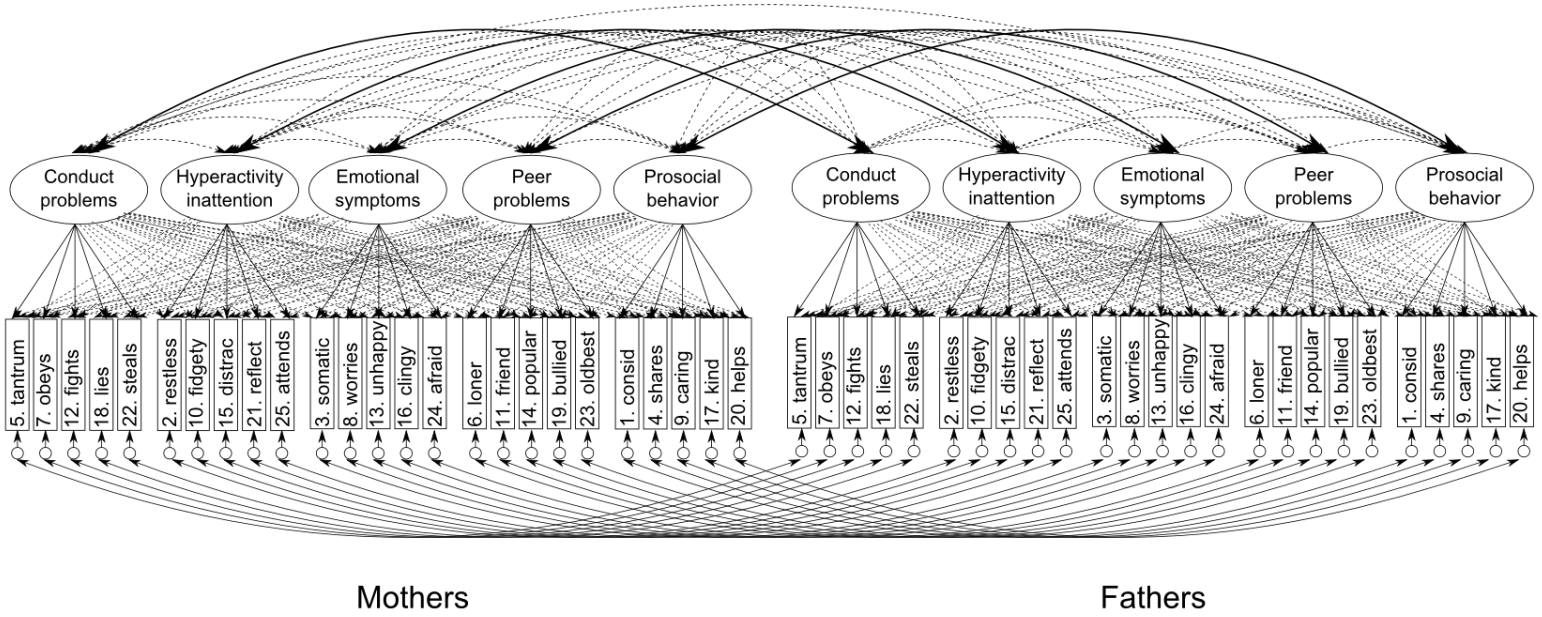
| | Mothers | | | | | Fathers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | COND | EMO | HYP | PEER | PRO | COND | EMO | HYP | PEER | PRO |
| Mothers | | | | | | | | | | |
| COND | | | | | | | | | | |
| EMO | *.23* | | | | | | | | | |
| HYP | *.36* | *.14* | | | | | | | | |
| PEER | .10 | *.25* | .04 | | | | | | | |
| PRO | *-.10* | *-.10* | *-.25* | -.07 | | | | | | |
| Fathers | | | | | | | | | | |
| COND | ***.73*** | .03 | *.24* | -.09 | *-.13* | | | | | |
| EMO | *.14* | ***.67*** | .07 | *.13* | *-.20* | *.28* | | | | |
| HYP | *.28* | .10 | ***.70*** | .09 | *-.10* | *.26* | *.13* | | | |
| PEER | .02 | .12 | .04 | ***.78*** | -.07 | .02 | *.24* | .08 | | |
| PRO | *-.20* | *-.12* | *-.17* | -.03 | ***.70*** | -.12 | *-.17* | *-.21* | -.11 | |

Note: . Italicized coefficients are significant at $p < .01$. Bolded coefficients are inter-parent agreement coefficients; COND=conduct problems; EMO= emotional symptoms; HYP= hyperactivity-inattention; PEER= peer problems; PRO= prosocial behavior.

**Figure caption**

Figure 1 Baseline model for the application of Exploratory Structural Equation Modeling to the invariance of the Strength and Difficulties Questionnaire measurement model between mothers and fathers.

Figure 1




Note: For ease of interpretation, full lines represent inter-parent agreement correlations and target loadings while dotted lines represent different factor-same source correlations (i.e., correlations between latent constructs within each parent), different-factor different source correlations (i.e., correlations between latent constructs between parents) and cross-loadings (i.e., loadings between a priori constructs and secondary items).

**Supporting Information for:**

**Evaluating Measurement Invariance between Raters using the Strength and Difficulties Questionnaire (SDQ)**

## 1. A review of journal articles analyzing the factor structure of the SDQ

Since 1999, there have been more than thirty studies examining the factor structure of the SDQ (see Table 1). This body of research is highly heterogeneous, since studies were deployed across 18 different countries, with sample sizes ranging from 128 to 71,840 participants, using parents', teachers', and individual self-reports. Moreover, these studies have differed in their factor analytic approach (Exploratory Factor Analysis [EFA], principal component analysis [PCA], confirmatory factor analysis [CFA]) and estimation methods (different kinds of maximum likelihood and weighted least squares). Attempts to replicate Goodman's original five-factor model of the SDQ (see Goodman, 2001) have yielded mixed results. Some studies have *supported* the five-factor model (e.g. d'Acremont & van Der Linden, 2008; Becker et al. 2004; Capron et al. 2007; Giannakopoulos et al. 2009; Hawes & Dadds, 2004; He et al. 2013; Hill & Hughes, 2007; Matsuishi et al. 2008; Niclasen et al. 2012; Rothenburger et al. 2008; Shevlin et al. 2012; Smedje et al. 1999; van Roy et al. 2008; Woerner et al. 2004; Yao et al. 2009), while others have reported *failed replications* (Dickey & Blumberg, 2004; Di Riso et al. 2010; Hagquist, 2007; Haynes et al. 2013; Muris et al. 2004). Alternative models have also been suggested such as those making a theoretically plausible distinction between prosocial, internalizing (merging COND and HYP) and externalizing (merging EMO and PEER) behaviors (Haynes et al., 2013; Goodman et al., 2010). Interestingly, the results of the comparison by Goodman (2010) were inconsistent across the three versions of the SDQ (parent, teacher, self-report): Parent data supported 3 factors, teacher data 5, and self-report both 3 and 5. Goodman and colleagues (2010) concluded that when a screen was sought for low-risk samples and populations, a 3 factor implementation of the SDQ would be appropriate, but when considering high-risk populations, then the original 5 factor model had noticeable benefits in terms of discriminant validity.

Table 1. Summary of studies since 1999 that have investigated factor structure in the Strengths and Difficulties Questionnaire

| *Year* | *Authors* | *Sample size* | *Age range* | *Country* | *Rater* | *Analytic strategy* | *Estimator* | *Results* |
|---|---|---|---|---|---|---|---|---|
| 1999 | Smedje et al. | 900 | 6-10 yrs | Sweden | Parents | PCA/Varimax | | 5-factor supported with cross-loadings |
| 2001 | Goodman | 10438 | 5-15 yrs | UK | Self, Parents, Teachers | EFA?/Varimax | | Expected 5-factor solutions with cross-loadings |
| 2001 | Koskelainen et al. | 1458 | 13-17 yrs | Finland | Self | EFA / Varimax | | 5-factor solutions with no simple structures, differences among boys and girls; 3-factor solution similar across gender with no simple structure |
| 2003 | Muris et al. | 562 | 9-15 yrs | Netherlands | Self, Parents | PCA/Oblimin | | 5-factor w/ cross loadings |
| 2004 | Muris et al. | 1111 | 8-13 yrs | Netherlands | Self | PCA/Oblimin | | 4-factor w/ cross loadings |
| 2004 | Becker et al. | 543 | 5-17 yrs | Germany | Parents, Teachers | CFA, PCA/Varimax | | CFA: AGFI=.85, RMR=.07, PCA: perfect solution |
| 2004 | Dickey & Blumberg | 10367 | 4-17 yrs | US | Parents/guardian | Cross validation PCA/PROMAX and CFA | ULS | Not very neat PCA 5-factor solution, better 3-factor solution, CFA used RMR and GFI |
| 2004 | Hawes & Dadds | 1359 | 4-9 yrs | Australia | Parent | PCA/Oblimin | | 5-factor supported, with cross-loadings |
| 2004 | Rønning et al. | 5225 | 11-16 yrs | Norway | Self | CFA | WLS | Poor fit 5-factor; added CUs and cross loadings |
| 2004 | Woerner et al. | 930 | 6-16 yrs | Germany | Parents | PCA/Varimax | | 5-factor supported with cross-loadings |
| 2005 | Kashala et al. | 1187 | 7-9 yrs | Congo | Teachers | PCA/Varimax | | 5-factor with no simple structure |
| 2006 | Van Leeuwen et al. | 1086 | 4-8 yrs | Netherlands | Parents, Teachers | PAF/oblique, CFA | | EFA 3- and 5-factor solution with cross-loadings; CFA 5-factor model slightly better than 3-factor model |

| *Year* | *Authors* | *Sample size* | *Age range* | *Country* | *Rater* | *Analytic strategy* | *Estimator* | *Results* |
|---|---|---|---|---|---|---|---|---|
| 2007 | Capron et al. | 1400 | 13 yrs | France | Self | PCA/Varimax | | Not very neat, substantial cross-loadings |
| 2007 | Hagquist | 8838 | 12-18 yrs | Sweden | Self | Rasch analysis | | |
| 2007 | Hill & Hughes | 784 | 6 yrs | US | Parents, Teachers | CFA | | 5-factor model marginally acceptable fit with CUs for oth parents and teachers |
| 2007 | Mellor & Stokes | 914 | 7-17 yrs | Australia | Self, Parents, Teachers | CFA | ML, ADF, MLR | Hierarchical (negative 2nd order factor) w/ poor fit |
| 2007 | Palmieri & Smith | 733 | 4-16 yrs | US | Grandmothers | CFA | MLR | (A) Hierarchical (negative 2nd order factor); (B) 5 factor model, (C) 5-factor w/ wording factor. All excellent fit but C better |
| 2008 | Rothernberger et al. | 2406 | 7-16 yrs | Germany | Self Parents | CFA, PCA/varimax | | 5-factor structure |
| 2008 | d'Acremont & Van der Linden | 560 | 12-14 yrs | Switzerland | Teachers | CFA | WLSMV, MLR | Only RMSEA and SRMR for single samples; CFI».80 for invariance |
| 2008 | Du et al. | 1965 | 3-17 yrs | China | Self (960), Parents, Teachers | PCA/Varimax | | 5-factor solutions with no simple structure in either rater |
| 2008 | Matsuishi et al. | 2899 | 4-12 yrs | Japan | Parents | EFA/Varimax | | 5-factor structure with no simple structure |
| 2008 | Percy et al. | 3753 | 12 yrs | Northern Ireland | Self | EFA/Promax | MLR | 3- and 5-factor poor fit event w/ CUs |
| 2008 | Ruchkin et al. | 4671 | 11-15 yrs | US | Self | CFA, PAF/Oblimin | | Original 5-factor partially supported, new 3-factor |
| 2008 | Van Roy et al. | 26269 | 10-19 yrs | Norway | Self, Parents (6645) | CFA | | (A) 5-factor model; (B) 5-factor model with CUs; (C) 5-factor model with wording factor; (C) batter fit; MTMM excellent fit |

| *Year* | *Authors* | *Sample size* | *Age range* | *Country* | *Rater* | *Analytic strategy* | *Estimator* | *Results* |
|---|---|---|---|---|---|---|---|---|
| 2009 | Giannakopoulos et al. | 1194 | 11-17 yrs | Greece | Self | CFA | ML | 5-factor model fitted after allowing secondary loadings |
| 2009 | Sanne et al. | 8999 | 7-9 yrs | Norway | Parents (6430), Teachers (8999) | EFA, ESEM, CFA, | WLSMV | Modestly modified version of original 5-factor, good support for informant invariance |
| 2009 | Yao et al. | 1135 | 11-18 yrs | China | Self | CFA | ML | original 5-factor + hierarchical, acceptable fit depending on age |
| 2010 | Di Riso et al | 1394 | 8-10 yrs | Italy | Self | CFA | WLS | 3-factor model slightly better fit than 5-factor |
| 2010 | Goodman A. et al. | 18222 | 5-16 yrs | UK | Self, Parents, Teachers | CFA | WLSMV | (A) 5-factor, (B) 5-factor w/ 2nd order, (C) 3-factor; B better model across informants but with CUs |
| 2010 | Mansbach-Kleinfeld et al. | 611 | 14-17 yrs | Israel | Self, Mothers | EFA/CFA | | Failed to replicate original 5-factor structure |
| 2010 | Stone et al. | Review | - | - | Parents, Teachers | | | Most studies confirmed 5-factor structure |
| 2011 | Richter et al. | 5379 | 15 yrs | Norway | Self | CFA | DWLS | Optimal fit 5-factor structure across ethnic groups |
| 2011 | Van de Looji-Jansen et al. | 11881 | 11-16 yrs | Netherlands | Self | EFA/CFA | WLSM | Original 5-factor model with CUs and new 4-factor model |
| 2012 | Essau et al. | 2418 | 12-17 yrs | Germany, Cyprus, England, Sweden, Italy | Self | CFA | MLR | Mixed results depending on country. Similar fit 3- and 5-factor on total sample, poor fit on national samples except Cyprus |
| 2012 | McCrory & Layte | 8514 | 9 yrs | Ireland | Parent | CFA | WLSMV | (A) 3-factor model, (B) 5 factor model, (C) 5-factor w/ wording factor, (D) Hierarchical (negative 2nd order factor); (C) better fitting |

| *Year* | *Authors* | *Sample size* | *Age range* | *Country* | *Rater* | *Analytic strategy* | *Estimator* | *Results* |
|---|---|---|---|---|---|---|---|---|
| 2012 | Niclasen et al. | 71840 | 5-12 yrs | Denmark | Parents, Teachers | PCA/Promax | | 5-factor w/ cross loadings |
| 2012 | Shevlin et al. | 202 | 7.17 yrs | Northern Ireland | Self, counsellor | CFA MTMM | WLSMV | Original 5-factor structure w/ cross-loadings to reach acceptable fit |
| 2013 | Gómez-Beneyto et al. | 6773 | 4-15 yrs | Spain | Informants | EFA (ULS, PROMIN), CFA | DWLS | 3- and 5-factor structure but no expected simple structure from EFA; adequate and similar fit for 3- and 5-factors |
| 2013 | Haynes et al. | 128 | 9-14 yrs | Australia | Self (with modified items) | PCA/Varimax and Oblimin | | 5-factor wth no simple structure; 3-factor but idiosyncratic solution (see Table 6) |
| 2013 | He et al. | 6843 | 13-18 yrs | US | Parents | CFA | WLSMV | Original 5-factor structure adequate |
| 2013 | Stone et al. | 1484 | 9-12 yrs | Netherlands | Mothers | CFA | WLSMV | 5-factor structure confirmed |
| 2013 | Niclasen et al. | 63615 | 5-7 yrs | Denmark | Parents, Teachers | CFA | WLS | 5-factor structure + second order factors |
| 2013 | Ezpeleta et al. | 622 | 3 yrs | Spain | Parents, Teachers | CFA | WLSMV | 5-factor structure confirmed |

Note: PCA: Principal Components Analysis; EFA = Exploratory Factor Analysys; CFA = Confrimatory Factor Analysis; ESEM = Exploratory Structural Equation Modeling; ULS = Unweighted Least Squares; WLS = Weighted Least Squares; ML = Maximum Likelihood; ADF = Asymptotic Distribution Free; MLR = robust Maximum Likelihood; WLSMV = Weighted Least Squares Mean and Variance Adjusted; DWLS = Diagonally Weighted Least Squares

## 2. Descriptive statistics for item and scale scores

Table 2 Descriptive statistics (proportions) for the Strengths and Difficulties Questionnaire items ($n$ = 695) on raw data (i.e., reverse item scores not reversed)

| | | | *Mothers* | | | | *Fathers* | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Item* | *Content* | *Scale* | *missing* | *0* | *1* | *2* | *missing* | *0* | *1* | *2* |
| sdq01 | consid | PRO | .01 | .01 | .42 | .56 | .01 | .02 | .52 | .45 |
| sdq02 | restles | HYP | <.01 | .51 | .40 | .09 | <.01 | .42 | .41 | .16 |
| sdq03 | somatic | EMO | .01 | .85 | .13 | .01 | <.01 | .86 | .12 | .02 |
| sdq04 | shares | PRO | <.01 | .03 | .58 | .39 | <.01 | .06 | .56 | .39 |
| sdq05 | tantrum | COND | <.01 | .51 | .39 | .10 | <.01 | .45 | .42 | .13 |
| sdq06 | loner | PEER | <.01 | .70 | .25 | .05 | .01 | .69 | .27 | .03 |
| sdq07 | obeys (r) | COND | <.01 | .05 | .60 | .36 | <.01 | .05 | .58 | .36 |
| sdq08 | worries | EMO | <.01 | .88 | .11 | .01 | <.01 | .89 | .09 | .01 |
| sdq09 | caring | PRO | <.01 | .02 | .29 | .69 | .01 | .01 | .28 | .70 |
| sdq10 | fidgety | HYP | .01 | .66 | .27 | .06 | .01 | .64 | .27 | .08 |
| sdq11 | friend (r) | PEER | <.01 | .02 | .12 | .86 | <.01 | .03 | .13 | .84 |
| sdq12 | fights | COND | <.01 | .90 | .09 | <.01 | <.01 | .89 | .10 | .01 |
| sdq13 | unhappy | EMO | <.01 | .92 | .06 | .02 | <.01 | .90 | .08 | .02 |
| sdq14 | popular (r) | PEER | <.01 | <.01 | .17 | .83 | .01 | .01 | .16 | .83 |
| sdq15 | distrac | HYP | <.01 | .47 | .45 | .07 | .01 | .43 | .45 | .10 |
| sdq16 | clingy | EMO | .01 | .51 | .41 | .08 | <.01 | .46 | .44 | .10 |
| sdq17 | kind | PRO | .01 | <.01 | .26 | .73 | .01 | .01 | .28 | .70 |
| sdq18 | lies | COND | <.01 | .82 | .17 | .01 | .02 | .78 | .19 | .01 |
| sdq19 | bullied | PEER | .01 | .91 | .07 | .01 | .01 | .92 | .07 | .01 |
| sdq20 | helpout | PRO | .01 | .03 | .45 | .51 | .01 | .05 | .48 | .46 |
| sdq21 | reflect (r) | HYP | .01 | .11 | .73 | .15 | .01 | .14 | .63 | .21 |
| sdq22 | steals | COND | <.01 | .96 | .04 | <.01 | <.01 | .96 | .03 | <.01 |
| sdq23 | oldbest | PEER | .01 | .72 | .23 | .04 | .01 | .73 | .22 | .03 |
| sdq24 | afraid | EMO | <.01 | .74 | .22 | .04 | <.01 | .69 | .27 | .04 |
| sdq25 | attends (r) | HYP | <.01 | .10 | .55 | .34 | .01 | .12 | .59 | .29 |

Note: (r) = reverse item; COND=conduct problems; EMO= emotional symptoms; HYP= hyperactivity-inattention; PEER= peer problems; PRO= prosocial behavior.

Table 3 Descriptive statistics for observed scale scores ($n$ = 695)

| *Scale* | *Valid* | *Min* | *Max* | *M* | *SD* | *SK* | *KU* |
|---|---|---|---|---|---|---|---|
| | | | | Mothers | | | |
| PRO | 680 | 2 | 10 | 7.79 | 1.67 | -0.57 | -0.24 |
| COND | 688 | 0 | 6 | 1.60 | 1.32 | 0.72 | 0.15 |
| EMO | 684 | 0 | 9 | 1.25 | 1.45 | 1.64 | 3.45 |
| HYP | 681 | 0 | 10 | 3.30 | 2.19 | 0.60 | -0.01 |
| PEER | 682 | 0 | 8 | 1.09 | 1.33 | 1.61 | 3.37 |
| Problems | 655 | 0 | 26 | 7.15 | 4.12 | 1.03 | 1.66 |
| | | | | Fathers | | | |
| PRO | 676 | 2 | 10 | 7.58 | 1.73 | -0.43 | -0.45 |
| COND | 678 | 0 | 8 | 1.73 | 1.42 | 0.90 | 0.87 |
| EMO | 688 | 0 | 9 | 1.37 | 1.43 | 1.41 | 2.71 |
| HYP | 675 | 0 | 10 | 3.59 | 2.30 | 0.41 | -0.33 |
| PEER | 673 | 0 | 8 | 1.09 | 1.34 | 1.53 | 2.61 |
| Problems | 648 | 0 | 26 | 7.80 | 4.33 | 0.95 | 1.58 |

Note: PRO= prosocial behavior; COND=conduct problems; EMO= emotional symptoms; HYP= hyperactivity-inattention; PEER= peer problems; Problems = total problem score is generated by summing the scores of the four problem subscales (excluding the prosocial behaviour subscale)

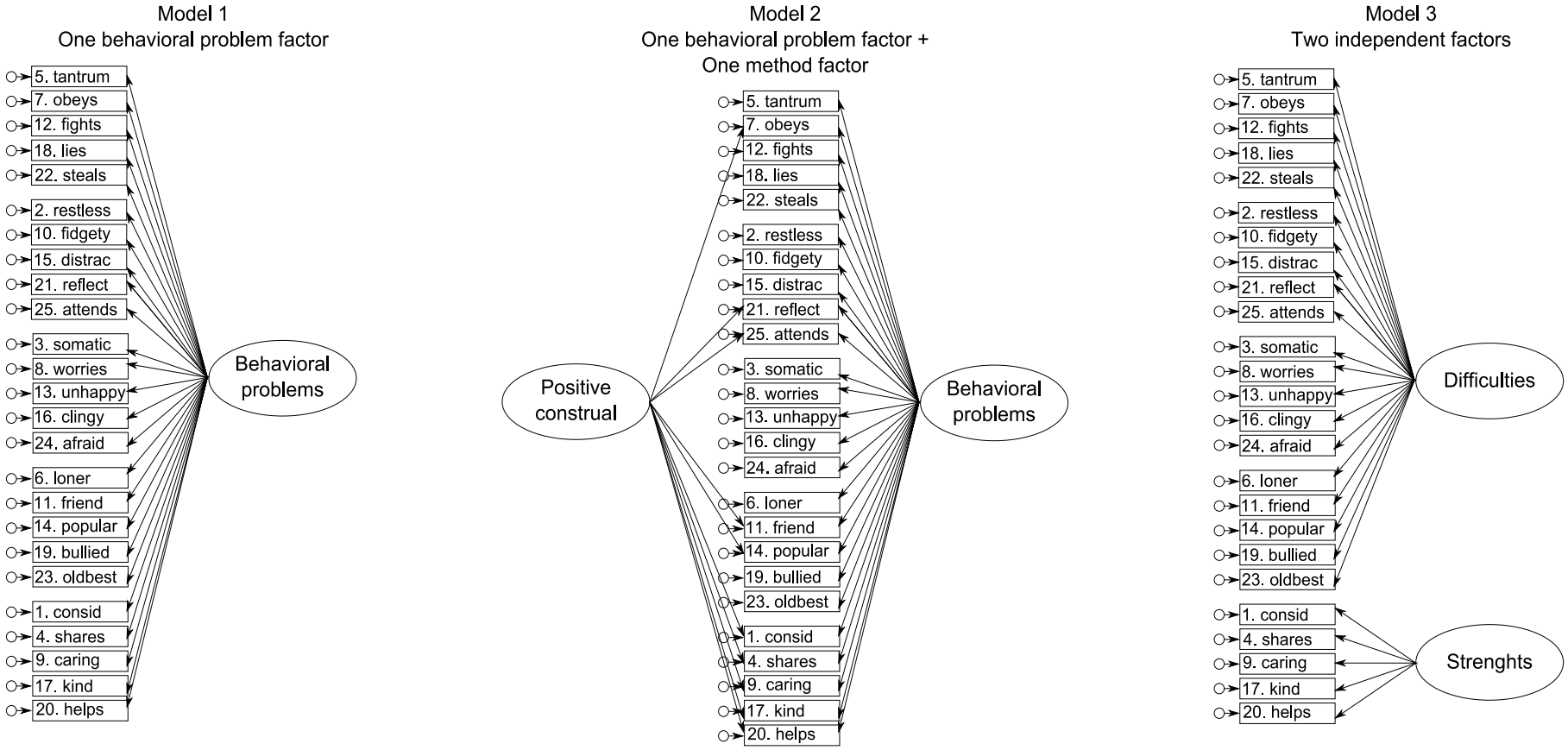
We also checked for abnormal and borderline score thresholds are for the total problem score, as reported in Stein et al. (2012): for parent reports, abnormal scores are 17+, borderline scores are 14-16. Details of the frequency distribution of total problems scores for mothers and fathers are reported in Table 4.

Table 4 Frequency distribution of total problems scores for mothers and fathers

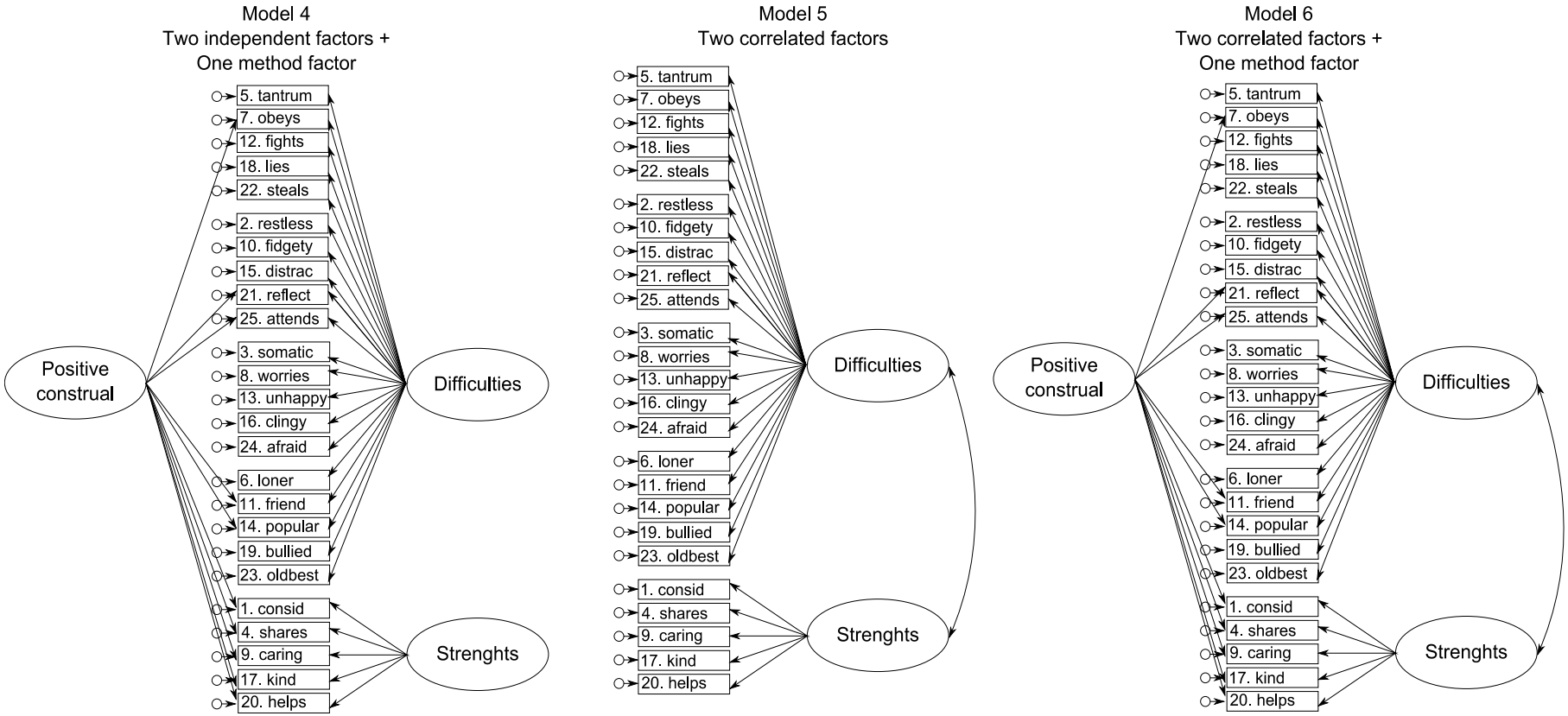
| *Classification of scores*[a] | *Score* | *Mothers* | | | | *Fathers* | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | *f* | *P* | *Valid P* | *Cumulative P* | *f* | *P* | *Valid P* | *Cumulative P* |
| Normal scores | 0 | 6 | .01 | .01 | .01 | 9 | .01 | .01 | .01 |
| | 1 | 24 | .03 | .04 | .05 | 17 | .02 | .03 | .04 |
| | 2 | 39 | .06 | .06 | .11 | 30 | .04 | .05 | .09 |
| | 3 | 47 | .07 | .07 | .18 | 36 | .05 | .06 | .14 |
| | 4 | 59 | .08 | .09 | .27 | 53 | .08 | .08 | .22 |
| | 5 | 79 | .11 | .12 | .39 | 62 | .09 | .10 | .32 |
| | 6 | 77 | .11 | .12 | .51 | 63 | .09 | .10 | .42 |
| | 7 | 60 | .09 | .09 | .60 | 67 | .10 | .10 | .52 |
| | 8 | 58 | .08 | .09 | .69 | 59 | .08 | .09 | .61 |
| | 9 | 51 | .07 | .08 | .76 | 62 | .09 | .10 | .71 |
| | 10 | 32 | .05 | .05 | .81 | 50 | .07 | .08 | .78 |
| | 11 | 41 | .06 | .06 | .87 | 33 | .05 | .05 | .83 |
| | 12 | 20 | .03 | .03 | .91 | 26 | .04 | .04 | .88 |
| | 13 | 12 | .02 | .02 | .92 | 24 | .03 | .04 | .91 |
| Borderline scores | 14 | 12 | .02 | .02 | .94 | 11 | .02 | .02 | .93 |
| | 15 | 10 | .01 | .02 | .96 | 11 | .02 | .02 | .95 |
| | 16 | 11 | .02 | .02 | .97 | 8 | .01 | .01 | .96 |
| Abnormal scores | 17 | 2 | .00 | .00 | .98 | 6 | .01 | .01 | .97 |
| | 18 | 1 | .00 | .00 | .98 | 4 | .01 | .01 | .97 |
| | 19 | 5 | .01 | .01 | .99 | 7 | .01 | .01 | .98 |
| | 20 | 2 | .00 | .00 | .99 | 2 | .00 | .00 | .99 |
| | 21 | 4 | .01 | .01 | 1.00 | 1 | .00 | .00 | .99 |
| | 22 | 1 | .00 | .00 | 1.00 | 2 | .00 | .00 | .99 |
| | 23 | 0 | .00 | .00 | 1.00 | 1 | .00 | .00 | .99 |
| | 24 | 0 | .00 | .00 | 1.00 | 1 | .00 | .00 | 1.00 |
| | 25 | 1 | .00 | .00 | 1.00 | 1 | .00 | .00 | 1.00 |
| | 26 | 1 | .00 | .00 | 1.00 | 2 | .00 | .00 | 1.00 |
| | Valid | 655 | .94 | | | 648 | .93 | | |
| | Missing | 40 | .06 | | | 47 | .07 | | |
| | Total | 695 | | | | 695 | | | |

Note: [a] as in Stein et al. (2012); *f* = observed frequency; *P* = proportion on total cases; *Valid P* = proportion on valid cases; *Cumulative P* = Cumulative proportion

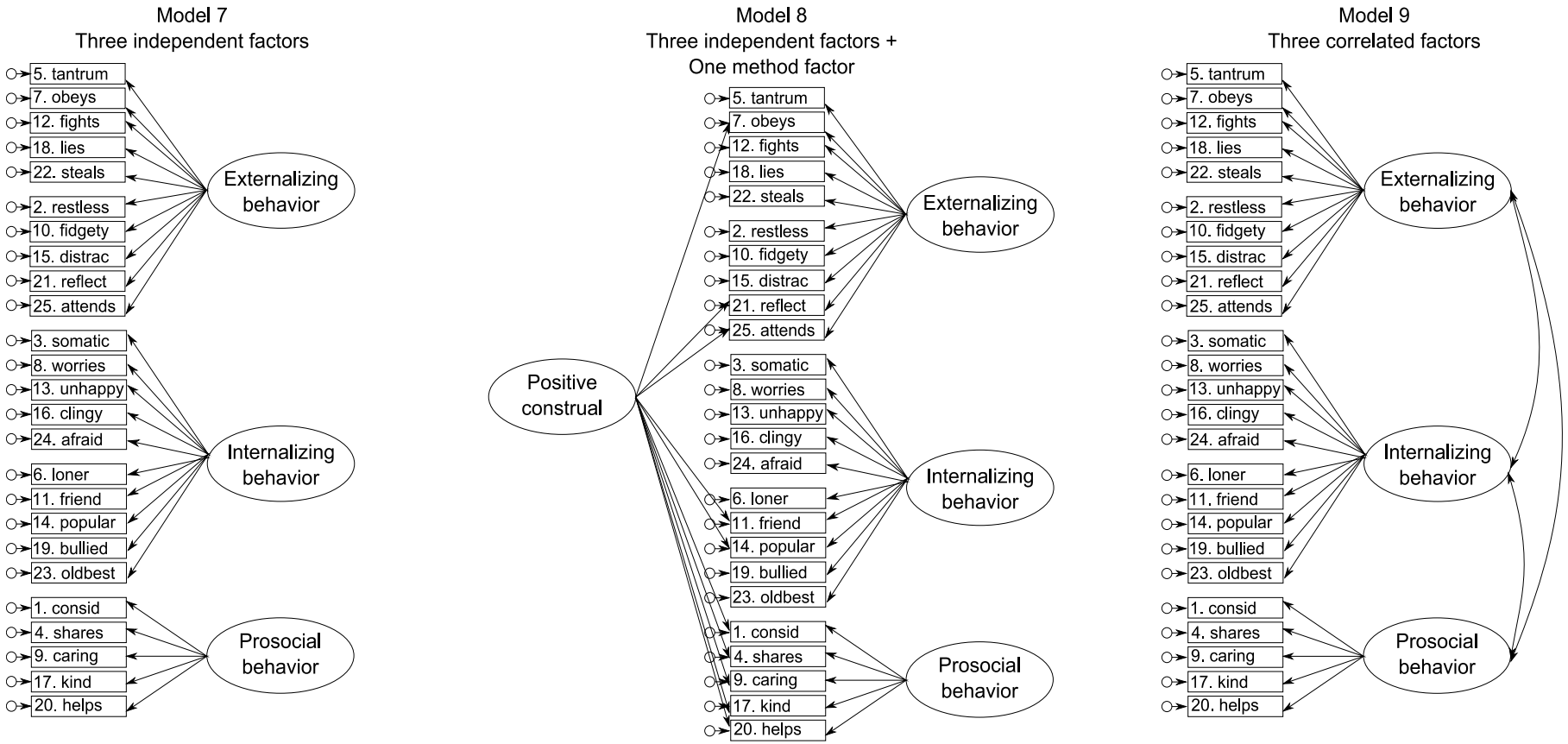
## 3. Confirmatory Factor Analysis models for the Strengths and Difficulties Questionnaire



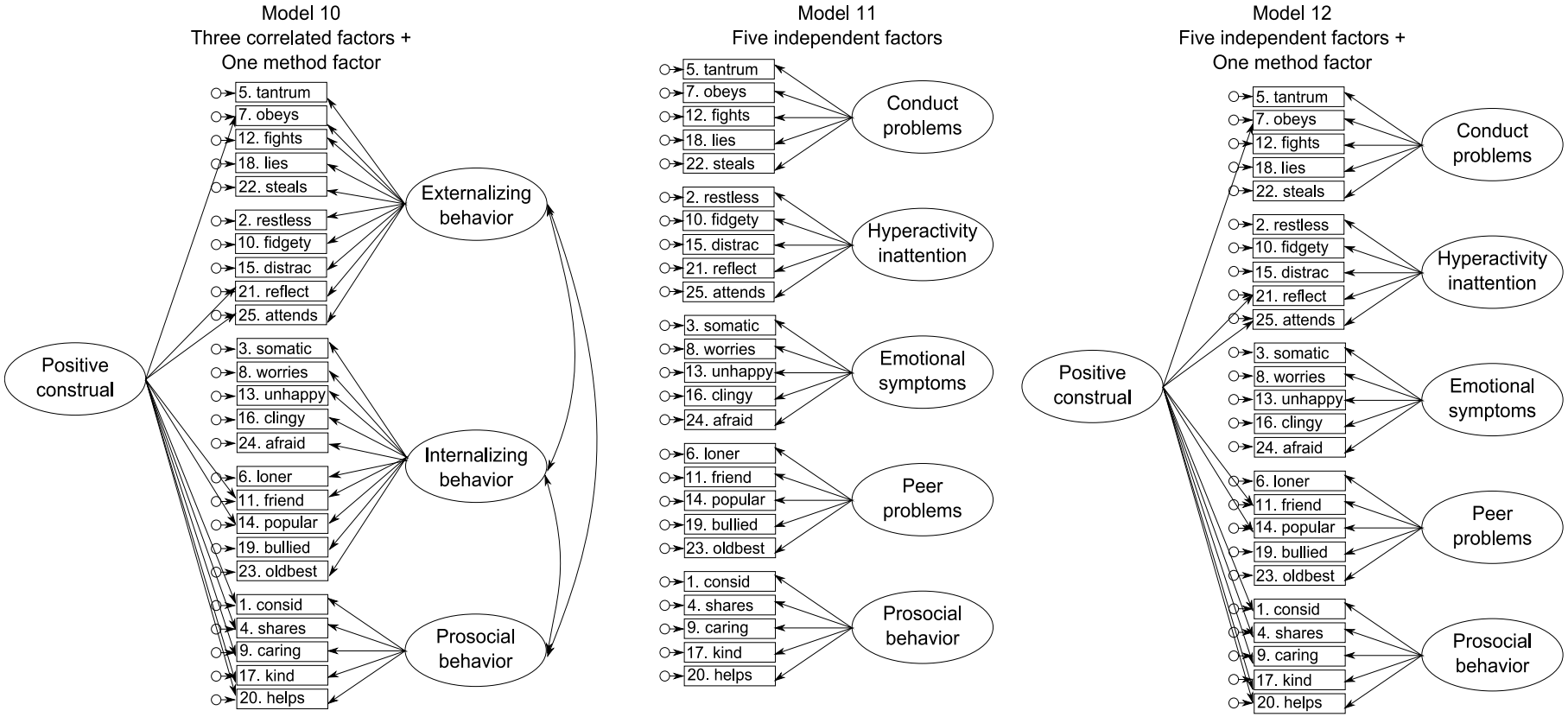
**3. Confirmatory Factor Analysis models for the Strengths and Difficulties Questionnaire (ctd.)**



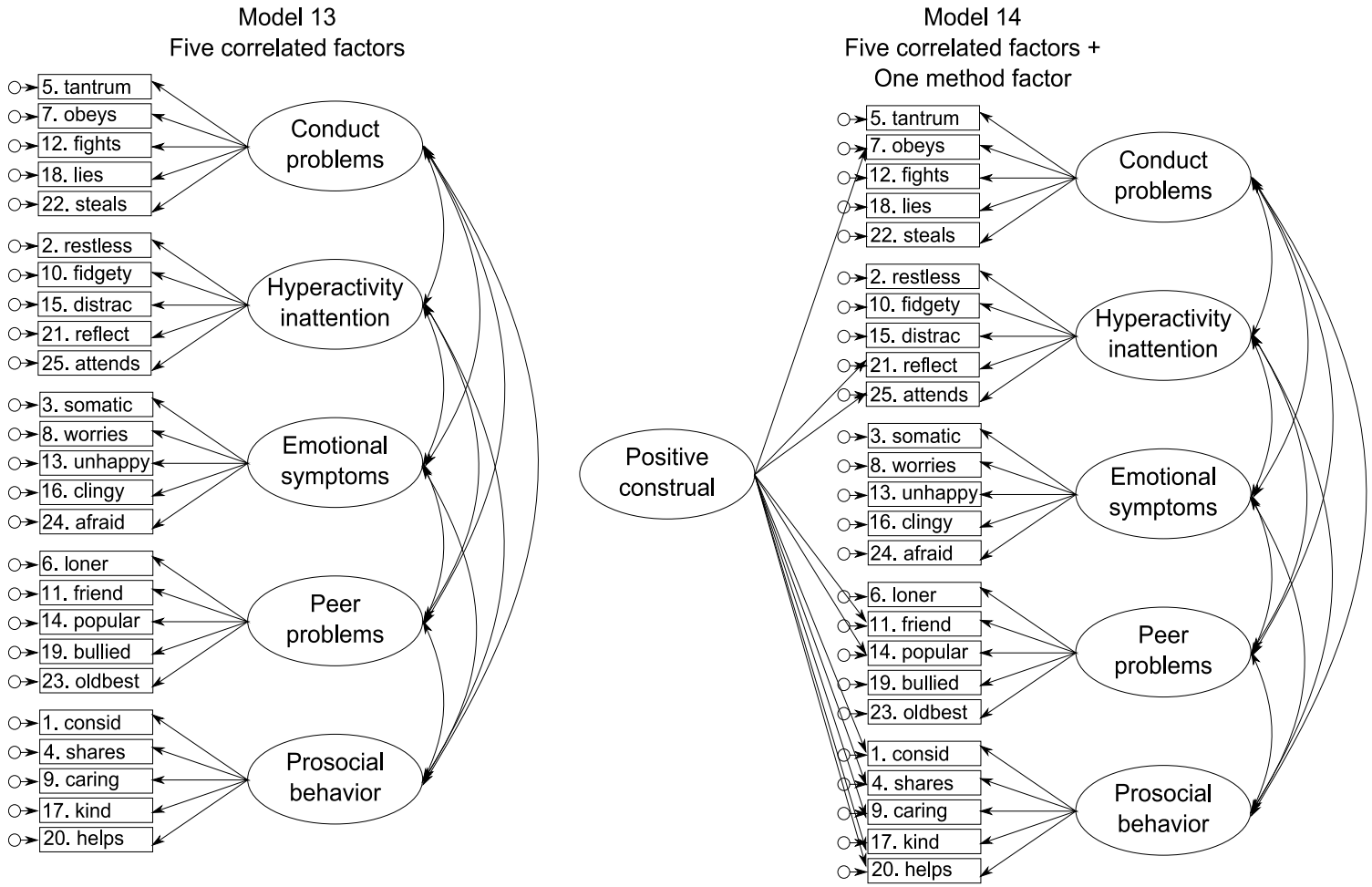
**3. Confirmatory Factor Analysis models for the Strengths and Difficulties Questionnaire (ctd.)**



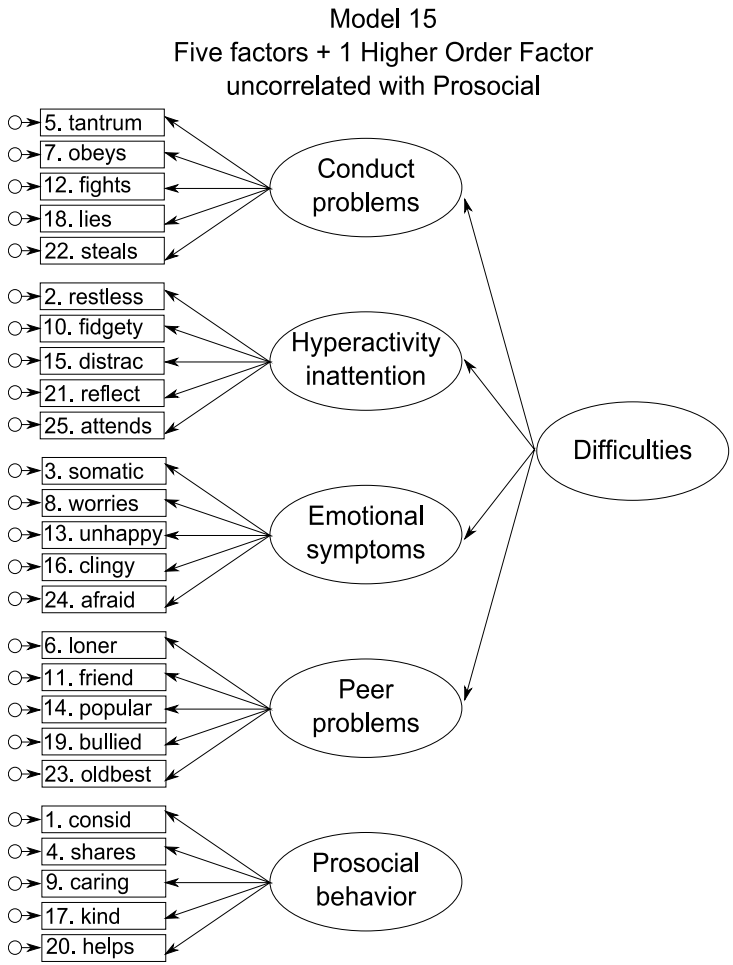
## 3. Confirmatory Factor Analysis models for the Strengths and Difficulties Questionnaire (ctd.)

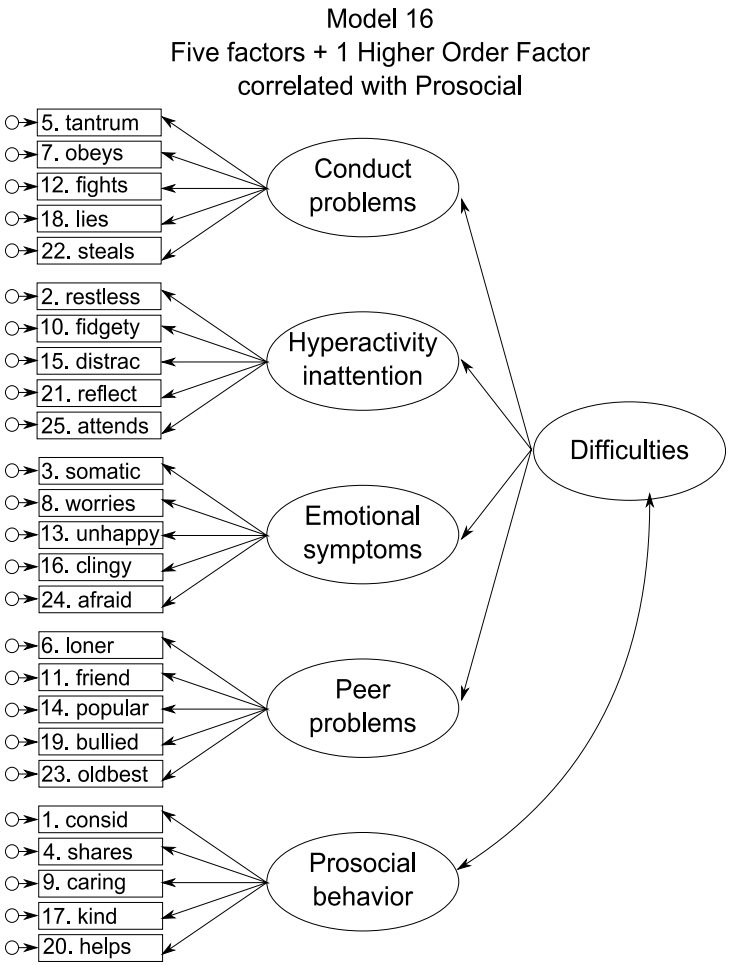


**3. Confirmatory Factor Analysis models for the Strengths and Difficulties Questionnaire (ctd.)**



**3. Confirmatory Factor Analysis models for the Strengths and Difficulties Questionnaire (ctd.)**



**4. Table 5 Results of Exploratory Structural Equation Models for fathers and mothers separately (n=695) standardized loadings**

| | | Fathers | | | | | | | | Mothers | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Item | Expected factor | 1 | 2 | 3 | 4 | 5 | $\tau_1$ | $\tau_2$ | RV | 1 | 2 | 3 | 4 | 5 | $\tau_1$ | $\tau_2$ | RV |
| 5. tantrum | COND | ***.85*** | ***.47*** | -.02 | -.18 | -.06 | -0.19 | 1.64 | .48 | ***.41*** | ***.55*** | -.02 | ***-.36*** | ***-.31*** | 0.03 | 1.78 | .53 |
| 7. obeys | COND | ***.62*** | .12 | .05 | -.18 | ***-.48*** | -0.46 | 2.15 | .56 | *.29* | .09 | .14 | -.19 | ***-.71*** | -0.50 | 2.28 | .55 |
| 12. fights | COND | ***.83*** | .09 | -.05 | .03 | -.13 | 1.66 | 3.17 | .57 | ***.71*** | .11 | .04 | .01 | *-.26* | 1.74 | 3.63 | .58 |
| 18. lies | COND | ***.52*** | *.27* | *.17* | -.05 | -.10 | 1.02 | 2.86 | .66 | .23 | ***.43*** | .14 | *-.30* | *-.24* | 1.15 | 3.13 | .65 |
| 22. steals | COND | ***.38*** | .20 | -.01 | -.17 | -.26 | 2.05 | 3.18 | .76 | .17 | ***.34*** | -.09 | -.13 | -.23 | 1.93 | 3.10 | .79 |
| 3. somatic | EMO | .15 | ***.47*** | .03 | -.07 | *.23* | 1.25 | 2.43 | .78 | .03 | ***.65*** | .01 | -.05 | .02 | 1.28 | 2.65 | .70 |
| 8. worries | EMO | .06 | ***.79*** | -.09 | ***.35*** | .14 | 1.69 | 3.01 | .53 | -.10 | ***1.01*** | -.08 | .12 | .04 | 1.67 | 3.58 | .50 |
| 13. unhappy | EMO | ***.46*** | ***.59*** | .05 | .27 | .05 | 1.82 | 2.99 | .52 | .07 | ***1.19*** | -.02 | -.01 | -.01 | 2.21 | 3.38 | .40 |
| 16. clingy | EMO | ***-.32*** | ***1.07*** | .00 | .02 | -.07 | -0.14 | 1.86 | .48 | ***-.51*** | ***1.04*** | .08 | .02 | .02 | 0.03 | 1.95 | .52 |
| 24. afraid | EMO | .02 | ***.75*** | .03 | .13 | .00 | 0.64 | 2.29 | .60 | -.19 | ***.93*** | -.02 | .10 | .03 | 0.85 | 2.39 | .56 |
| 2. restles | HYP | ***.85*** | -.11 | ***1.01*** | .06 | .01 | -0.33 | 1.73 | .33 | ***1.33*** | .00 | ***1.13*** | .02 | -.02 | 0.05 | 2.87 | .22 |
| 10. fidgety | HYP | ***.73*** | .01 | ***.82*** | .07 | .10 | 0.59 | 2.22 | .40 | ***1.12*** | -.02 | ***.94*** | .02 | .12 | 0.78 | 2.89 | .30 |
| 15. distrac | HYP | .20 | .14 | ***1.41*** | .04 | .00 | -0.28 | 2.28 | .30 | .24 | .15 | ***1.40*** | .01 | .00 | -0.12 | 2.59 | .31 |
| 21. reflect | HYP | .13 | *-.19* | ***.64*** | -.10 | ***-.52*** | -1.04 | 1.41 | .57 | -.02 | -.01 | ***.48*** | *-.16* | ***-.42*** | -1.26 | 1.51 | .65 |
| 25. attends | HYP | -.05 | .01 | ***1.22*** | *-.27* | ***-.56*** | -0.93 | 1.99 | .35 | -.03 | -.02 | ***1.85*** | -.03 | ***-.48*** | -0.91 | 2.85 | .20 |
| 6. loner | PEER | -.13 | ***.33*** | .11 | ***.57*** | *-.19* | 0.64 | 2.40 | .63 | .03 | ***.33*** | .05 | ***.57*** | -.03 | 0.66 | 2.05 | .65 |
| 11. friend | PEER | .08 | -.02 | -.05 | ***.45*** | ***-.47*** | 1.21 | 2.25 | .70 | .00 | .02 | .01 | ***.71*** | ***-.52*** | 1.43 | 2.80 | .57 |
| 14. popular | PEER | -.03 | .11 | -.11 | ***.49*** | ***-.71*** | 1.29 | 3.25 | .57 | .05 | .16 | .14 | ***.52*** | ***-.52*** | 1.23 | 3.86 | .60 |
| 19. bullied | PEER | .15 | .06 | .03 | ***.44*** | *.24* | 1.61 | 2.66 | .76 | .15 | ***.32*** | -.03 | ***.41*** | -.06 | 1.62 | 2.62 | .72 |
| 23. oldbest | PEER | .00 | .03 | .06 | ***1.88*** | -.02 | 1.38 | 3.92 | .22 | ***.41*** | .16 | -.02 | ***.67*** | .01 | 0.80 | 2.23 | .58 |
| 1. consid | PRO | ***-.45*** | -.01 | -.09 | -.01 | ***.87*** | -2.90 | 0.16 | .47 | -.14 | .01 | .05 | .02 | ***1.09*** | -3.56 | -0.25 | .45 |
| 4. shares | PRO | -.16 | -.12 | -.04 | -.10 | ***.58*** | -1.93 | 0.35 | .68 | -.04 | -.07 | -.09 | -.04 | ***.56*** | -2.15 | 0.34 | .73 |
| 9. caring | PRO | .00 | .04 | .05 | -.11 | ***.82*** | -2.93 | -0.70 | .60 | .10 | .13 | .09 | *-.29* | ***1.08*** | -3.05 | -0.75 | .46 |
| 17. kind | PRO | ***-.33*** | .07 | *.23* | ***-.32*** | ***.86*** | -3.37 | -0.76 | .50 | .01 | .12 | .00 | ***-.32*** | ***1.05*** | -3.83 | -0.90 | .47 |
| 20. helpout | PRO | .20 | -.07 | -.15 | .01 | ***.91*** | -2.22 | 0.11 | .54 | ***.37*** | -.05 | -.13 | -.04 | ***.70*** | -2.31 | -0.03 | .63 |
| | *r* with 2 | *.26* | | | | | | | | *.40* | | | | | | | |
| | *r* with 3 | *.23* | *.19* | | | | | | | .18 | *.16* | | | | | | |
| | *r* with 4 | *.18* | *.30* | *.16* | | | | | | -.02 | .21 | -.08 | | | | | |
| | *r* with 5 | *-.16* | -.15 | -.06 | .00 | | | | | -.16 | *-.16* | *-.24* | .03 | | | | |

Note: Bolded coefficients are higher than |.30|. Italicized coefficients are significant at p <. 01. $\tau_1$ and $\tau_2$: item thresholds; RV = items residual variance; COND=conduct problems; EMO= emotional symptoms; HYP= hyperactivity-inattention; PEER= peer problems; PRO= prosocial behavior

**References for this Supporting Information**

Becker, A., Woerner, W., Hasselhorn, M., Banaschewski, T., & Rothenberger, A. (2004). Validation of the parent and teacher SDQ in a clinical sample. *European Child & Adolescent Psychiatry, 13 [Suppl 2]*, II/11-II/16. doi: 10.1007/s00787-004-2003-5.

Capron, C., Therond, 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.

d'Acremont, M., & Van der Linden, M. (2008). Confirmatory factor analysis of the Strengths and Difficulties Questionnaire in a community sample of French-speaking adolescents. *European Journal of Psychological Assessment., 24,* 1-8. doi: 10.1027/1015-5759.24.1.1.

Di Riso, D., Salcuni, S., Chessa, D., Raudina, A., Lis, A., & Altoé, G. (2010). The Strengths and Difficulties Questionnaire (SDQ). Early evidence of its reliability and validity in a community sample of Italian children. *Personality and Individual Differences, 49,* 570-575. doi: 10.1016/j.paid.2010.05.005.

Dickey, W., C., & Blumberg, S., J. (2004). Revisiting the factor structure of the Strengths and Difficulties Questionnaire: United States, 2001. *Journal of the American Academy of Child & Adolescent Psychiatry, 43,* 1159-1167. doi: 10.1097/01.chi.0000132808.36708.a9.

Du, Y., Kou, J., & Coghill, D. (2008). The validity, reliability and normative scored 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.

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*, 232-245. doi: 10.1002/mpr.1364.

Ezpeleta, L., Granero, R., de la Osa, N., Penelo, E., & Domenèch, J. M. (2013). Psychometric properties of the Strengths and Difficulties Questionnaire 3–4 in 3-year-old preschoolers. *Comprehensive Psychiatry*, *54*, 282-291. doi: 10.1016/j.comppsych.2012.07.009.

Giannakopoulous, 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. doi:10.1186/1744-859X-8-2.

Gomez-Beneyto, M., Nolasco, A., Moncho, J., Pereyra-Zamora, P., Tamayo-Fonseca, N., Munarriz, M., Salazar, J., Tabarés-Seisdedos, R., & Girón, M. (2013). Psychometric behaviour of the Strengths and Difficulties Questionnaire (SDQ) in the Spanish National Health Survey 2006. *BMC Psychiatry*, *13*, 95. doi: 10.1186/1471-244X-13-95.

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*, 1179-1191. doi: 10.1007/s10802-010-9434-x.

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

Hagquist, C. (2007). The psychometric properties of the self-reported SDQ – An analysis of Swedish data based on the Rasch model. *Personality and Individual Differences*, *43*, 1289-1301. doi: 10.1016/j.paid.2007.03.022

Hawes, D., J., & Dadds, M. R. (2004). Australian data and psychometric properties of the Strengths and Difficulties Questionnaire. *Australian & New Zealand Journal of Psychiatry, 38*, 644-651. doi: 10.1111/j.1440-1614.2004.01427.x.

Haynes, A., Gilmore, L., Shochet, I., Campbell, M., & Roberts, C. (2013). Factor analysis of the self-report version of the strengths and difficulties questionnaire in a sample of children with intellectual disability. *Research In Developmental Disabilities, 34*, 847-854. doi: 10.1016/j.ridd.2012.11.008.

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

Hill, C., R., & Hughes, J. N. (2007). An examination of the convergent and discriminant validity of the Strengths and Difficulties Questionnaire. *School Psychology Quarterly, 22*, 380-406. doi: 10.1037/1045-3830.22.3.380.

Kashala, E., Elgen, I., Sommerfelt, K. & Tylleskar, T. (2005). Teacher ratings of mental health among school children in Kinshasa, Democratic Republic of Congo. *European Child & Adolescent Psychiatry. 14*, 208-215. doi: 10.1007/s00787-005-0446-y.

Koskelainen, M., Sourander, A., & Kaljonen, A. (2000). The Strengths and Difficulties Questionnaire among Finnish school-aged children and adolescents. *European Child & Adolescent Psychiatry, 9*, 277-284. doi: 10.1007/s007870070031.

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. *Frontiers in Psychiatry, 1*, 151. doi: 10.3389/fpsyt.2010.00151.

Matsuishi, T., Nagano, M., Araki, Y., Tanaka, Y., Iwasaka, M., Yamashita, Y., Nagamitsu, S., Lizuka, C., Ohya, T., Shibuya, K., Hara, M., Matsuda, K., Tsuda, A., & Kakuma, T. (2008). Scale properties of the Japanese version of the Strengths and Difficulties Questionnaire (SDQ): A study of infant and school children in community samples. *Brain & Development, 30,* 410-415. doi: 10.1016/j.braindev.2007.12.003.

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.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale NJ: Erlbaum.

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

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 & Adolescent Psychiatry*, 12, 1-8. doi: 10.1007/s00787-003-0298-2.

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. *British Journal of Clinical Psychology*, *43,* 437-448. doi: 10.1348/0144665042388982.

Niclasen, J., Skovgaard, A. M., Nybo Andersen, A.-M., Sømhovd, M. J., & Obel, C. (2012). 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.

Niclasen, J., Teasdale, T. W., Nybo Andersen, A.-M., Skovgaard, A. M., Elberling, H., & Obel, C. (2012). Psychometric properties of the Danish Strength and Difficulties Questionnaire: The SDQ assessed for more than 70,000 raters in four different cohorts. *PLOS One*, e32025. doi: 10.1371/journal.pone.0032025.

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.

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.

Raykov, T. (1997). Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research, 32*, 329-353. doi:10.1207/s15327906mbr3204_2.

Richter, J., Sagatun, A., Heyerdahl, S., Oppedal, B., & Roysamb, 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*, *52*, 1002-1011. doi: 10.1111/j.1469-7610.2011.02372.x.

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

Rothenberger, A., Becker, A., Erhart, M., Wille, N., Ravens-Sieberer, U., & the BELLA Study Group (2008). Psychometric properties of the parent strengths and difficulties questionnaire in the general population of German children and adolescents: results of the BELLA study. *European Child & Adolescent Psychiatry, 17 [Suppl 1],* 99-105. doi: 10.1007/s00787-008-1011-2.

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.

Sanne, B., Torsheim, T., Heiervang, E., & Stormark, K. M. (2009). The Strengths and Difficulties Questionnaire in the Bergen Child Study: A conceptually and methodically motivated structural analysis. *Psychological Assessment, 21,* 352-364. doi: 10.1037/a0016317.

Shevlin, M., Murphy, S., McElearney, A., O'Kane, D., Tracey, A., & Adamson, G. (2012). Confirmatory factor analysis of adolescent self- and informant ratings of the Strengths and Difficulties Questionnaire. *Irish Journal of Psychology, 3,* 17-28. doi: 10.1080/03033910.2011.649569.

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

Stein, A., Malmberg, L-E., Sylva, K., Leach, P., Barnes, J., & the FCCC group (2012). The influence of different forms of early childcare on children's emotional and behavioural development at school entry. *Child: Care, Health and Development, 39,* 676-687. doi: 10.1111/j.1365-2214.2012.01421.x.

Stone, L. L., Otten, R., Engels, R. C., Vermulst, Ad A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: A review. *Clinical Child and Family Psychology Review*, *13*, 254-274. doi: 10.1007/s10567-010-0071-2.

Stone, L. L., Ottne, R., Ringlever, L., Hiemstra, M., Engels, R. C. M. E., Vermulst, Ad A., & Janssens, J. M. A. 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.

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*, 127-144. doi: 10.1348/014466510X498174.

Van Leeuwen, K., Meerschaert, T., Bosmans, G., De Medts, L., & Braet, C. (2006).The Strengths and Difficulties Questionnaire in a community sample of young children in Flanders. *European Journal of Psychological Assessment., 22,* 189-197. doi: 10.1027/1015-5759.22.3.189.

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.

Woerner, W., Becker, A., & Rothenberger, A. (2004). Normative data and scale properties of the German parent SDQ. *European Child & Adolescent Psychiatry*, 13, 3-10. doi: 10.1007/s00787-004-2002-6.

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*, 55-62. doi: 10.1016/j.jadohealth.2008.11.006.