Do They Rate in the Same Way?

Testing of Measurement Invariance Across Parent and Teacher SDQ Ratings

Jana Rogge,1,2 Ute Koglin,1 and Franz Petermann3

1Department of Special Education and Rehabilitation, C. v. Ossietzky University Oldenburg, Germany
2Department of Pedagogical and Educational Sciences, University of Groningen, The Netherlands
3Center of Clinical Psychology and Rehabilitation, University of Bremen, Germany

Abstract: The internal structure of the Strengths and Difficulties Questionnaire (SDQ) has been widely discussed in the literature. The first purpose of this study was, therefore, to analyze the internal structure of the SDQ as completed by German parents and teachers. Parents and teachers of 1,135 children between the ages of 3 and 6 years in German daycare facilities were asked to complete the SDQ. Using confirmatory factor analysis (CFA), we found that the original five-factor model with added correlations between items with similar content and allowing cross-loads of reverse-worded items to the prosocial behavior factor is the model that fits the data the best. Additionally, we found significant mean differences between parents and teachers in four of the five SDQ subscales. Nevertheless, to compare the scores across both groups, it is necessary that the SDQ measures social-emotional and behavioral difficulties for both parents and teachers in the same way. Hence, the second purpose of this study was to test for measurement invariance of the SDQ across both groups. Using multigroup confirmatory factor analysis (MGCFA), we found that strict measurement invariance is tenable. Therefore, the observed means of the SDQ scores from parents and teachers are readily comparable and interpretable.

Keywords: SDQ, measurement invariance, factorial structure, confirmatory factor analysis, multigroup confirmatory factor analysis

A recent study on the prevalence of psychosocial problems found that almost 18% of German children and adolescents are affected by emotional and/or behavioral problems (Hölling, Schlack, Petermann, Ravens-Sieberer, & Mauz, 2014). An early identification of these problems is important to prevent them from developing into more severe psychiatric disorders (Stone, Otten, Engels, Vermulst, & Janssens, 2010). Since psychosocial problems are highly context specific, it is important to obtain information about social-emotional and behavioral difficulties from multiple adults who know the child in different social contexts (Achenbach, 1991; Goodman, Ford, Simmons, Gatward, & Meltzer, 2000; Stone et al., 2010).

The Strengths and Difficulties Questionnaire (SDQ) is a widely used screening instrument to capture behavioral and emotional problems in children and adolescents of ages 3-16 years. The SDQ evaluates the responses from parents and teachers of those children, as well as the children themselves (Goodman, 1997). Since parents and teachers observe children under different circumstances where diverse behaviors are shown, the SDQ provides an important measurement to capture information about children’s social-emotional and behavioral difficulties. The SDQ is composed of 25 items that represent five subscales: hyperactivity, emotional symptoms, conduct problems, peer problems, and prosocial behavior (Goodman, 1997; see also Appendix).

Psychometric Quality of the SDQ

Due to its widespread application, the psychometric quality of the SDQ has been examined in numerous studies. Regarding the internal consistency of the SDQ scales, Van de Looij-Jansen, Goedhart, de Wilde, and Treffers (2011) pointed out that several studies have found low values of coefficient α for one or more subscales of the self-reported form of the SDQ. Additionally, Kersten et al. (2015) identified that Cronbach’s α with a weighted average between 0.49 and 0.69 for parent subscales falls below the acceptable threshold of 0.70, while almost all values for Cronbach’s α for the teacher version reach this threshold. Therefore, it seems that the teacher version has a higher internal consistency than the parent and self-reported versions.
With respect to discriminant validity, Goodman, Lamping, and Ploubidis (2010) indicated that the behavior, hyperactivity, and prosocial subscales show especially poor discriminant validity. Moreover, Hill and Hughes (2007) concluded that the SDQ does not appear to discriminate well among the five latent constructs.

On the other hand, discriminant validity has been confirmed in many other studies. For instance, Becker, Woerner, Hasselhorn, Banaschewski, and Rothenberger (2004) found that the SDQ for German parents and teachers is useful for discriminating between a community and a clinical sample of children aged 5–17 years. In addition, Petermann, Petermann, and Schreyer (2010) have shown that the SDQ for teachers (especially the peer problems subscale) is suitable for differentiating between preschoolers with and without developmental delay.

With regard to the convergent validity of the SDQ, Kersten et al. (2015) found a varying picture concerning the five subscales. Using weighted average correlation coefficients for 16 studies, they concluded that convergent validity is given for three subscales (emotional problems, conduct problems, and hyperactivity) and the total difficulties problem scale for both the parent and teacher versions of the SDQ. However, the subscales for neither prosocial behavior nor peer problems reach acceptable correlations.

By comparing SDQ scores with the findings of clinical diagnostic interviews, Bekker, Bruck, and Seiberras (2013), Goodman (2001), Ezpeleta, Grancro, de la Osa, Penelo, and Domenec (2012), and Mathai, Anderson, and Bourne (2004) found inadequate support for criterion validity. These findings are in contrast to those of Goodman, Renfrew, and Mullick (2000) who indicated high sensitivities for scores of the three subscales of peer problems, conduct problems, and hyperactivity.

With respect to the factor structure (structural validity) of the parent version of the SDQ, Goodman et al. (2010) found that the use of the broader three-factor structure (externalizing problems, internalizing problems, and prosocial behavior) allows for a better solution than the original five-factor model. When analyzing the internal structure of the self-reported SDQ, Di Riso et al. (2010), Dickey and Blumberg (2004), and Percy, McCristal, and Higgins (2008) came to a similar conclusion. However, Ruchkin, Jones, Vermeiren, and Schwab-Stone (2008), Ruchkin, Koposov, and Schwab-Stone (2007), Zhang, Jing, McWhinnie, and Abela (2009), Burstein and Schmitz (2012), Svedin and Pribe (2008), and Van Roy, Veenstra, and Clench-Aas (2008) have shown that the five-factor solution fits the best with self-reported SDQ scores. In addition, Essau et al. (2012) concluded that the factor structure of the self-reported SDQ version differs across countries. However, neither the three- nor five-factor structure achieved an acceptable model fit in Germany, Sweden, Italy, and the United Kingdom. In contrast, both the three- and five-factor solutions reached an adequate model fit in Cyprus (Essau et al., 2012). Using the SDQ teacher questionnaire, Niclasen, Skovgaard, Andersen, Somhovd, and Obel (2012) and Van Leeuwen, Meerschaert, Bosmans, De Medts, and Braet (2006) indicated a good fit for the five-factor solution. Interestingly, in neither study did the parent SDQ version reach an acceptable model fit for the five-factor solution. Hill and Hughes (2007), Mieloo et al. (2012), and Rothenberger et al. (2008) concluded that fit indices for the five-factor solution only reach a marginally adequate fit for parent and teacher SDQ scores.

These findings indicate that the internal structure of the SDQ varies depending on the respective SDQ version and the sample characteristics (country, age, etc.). Therefore, the first purpose of this study was to analyze the internal structure of parent and teacher SDQ in Germany.

**Measurement Invariance**

Empirical studies have also found significant differences across parent and teacher SDQ scores for all subscales (Becker et al., 2004; Cefai, Camilleri, Cooper, & Saíd, 2011; Mieloo et al., 2012; Stone et al., 2010). Based on these findings, the question becomes whether the differences really reflect the dissimilar evaluation context of parents and teachers or whether the findings are the result of measurement variance across the parent and teacher groups. Therefore, before comparing the parent and teacher SDQ scores, it has to be shown that the underlying latent dimensions of the SDQ, measuring the same construct in the same way, and that the measurements themselves operate in the same way across parents and teachers.

The presence of measurement invariance is a necessary condition when comparing and interpreting scores between different groups (Davidov, 2010). A lack of measurement invariance between groups can lead to two problems (Chen, 2008). Firstly, the same measurement instrument may measure diverse constructs across different groups. Secondly, even if the same instrument measures the same construct, the construct might have different meanings for different groups (Cieciuch, Davidov, Oberski, & Algesheimer, 2015; Davidov, 2010). In both cases it might be that differences observed between groups do not correspond with real differences across groups, and therefore are likely to be invalid.

Depending on the severity of model restrictions, different levels of invariance are distinguishable (Brown, 2006). Configural invariance represents the lowest level and is indicated if factor structures are the same across groups. If the assumption of configural invariance holds,
the instrument is expected to measure the same construct across different groups. To assume that a construct has the same meaning across different groups (metric invariance), factor structure and factor loadings have to be equal in all groups. For straightforward interpretation of latent means and correlations across groups, both factor loadings and intercepts (or, in the case of ordinal indicators, thresholds) must be the same across groups (scalar invariance; Acoc, 2013; Davidov, 2010). For a meaningful comparison of observed means, in addition to loading and thresholds, the residual variance for all items has to be equal across groups (strict measurement invariance; Wu et al., 2007). Therefore, to compare the observed means of the SDQ subscales, strict measurement invariance across parent and teacher SDQ ratings must exist.

Analysis of the SDQ scores between parents and teachers is usually carried out with the assumption that strict measurement invariance of the measures between both groups exists (Chiorri, Hall, Casely-Hayford, & Malmberg, 2015; Goodman, 1997). Since the existence of strict measurement invariance determines the means of assessing evaluation differences between parents and teachers, knowledge of whether measurement invariance exists between the groups has far-reaching significance for both research and praxis. However, to the best of our knowledge no study has checked for measurement invariance across the evaluation of German parents and teachers based on the SDQ. Therefore, the second purpose of this study was to check for strict measurement invariance between German parent and teacher SDQ scores.

### Methods

#### Sample and Procedure

Using the parent and teacher SDQ questionnaires, data was collected for 1,135 children aged 3–6 years. These children attended daycare facilities in Lower Saxony, Bremen, Hamburg, and Hesse. In total, we received complete evaluations for 1,057 children (52.4% female). The children were, on average, 4.6 years old ($SD = 0.98$) years. Unfortunately, we did not obtain any information on whether the father or mother of the child completed the SDQ. Additionally, we do not have any other background information about the characteristics of the children and parents.

#### Instrument

We used the SDQ for parents and teachers. The SDQ is an efficient and freely available questionnaire used to assess psychological difficulties in children and adolescents between the ages of 3 and 16 years (Goodman, Ford, et al., 2000). The SDQ is composed of 25 items that are divided among five subscales: hyperactivity, emotional symptoms, conduct problems, peer problems, and prosocial behavior (Goodman, 1997). Each item implies a statement where the evaluator (parent or teacher) chooses among three response options ($0 = \text{not true}$, $1 = \text{somewhat true}$, $2 = \text{certainly true}$). The score in all five subscales can therefore range from 0 to 10. In addition, the scores for all subscales except prosocial behavior can be summed to create a total difficulties score that ranges from 0 to 40 (Goodman, 1997).

In concurrence with previous studies (Cefai et al., 2011; Kersten et al., 2015; Stone et al., 2010), we found low coefficient $\alpha$ values for four of the five subscales in the SDQ parent version (Table 1). In contrast, the teachers' values for Cronbach's $\alpha$ range from acceptable ($\geq 0.70$) to good ($\geq 0.80$). Therefore, it seems that the teacher version has a higher internal consistency than the parent and self-reported versions.

We have calculated the descriptive statistics of the SDQ subscales and the total difficulties score for parents, teachers, and the whole sample. The results are displayed in Table 2. In accordance with Mieloo et al. (2012), Stone et al. (2010), Cefai et al. (2011), and Becker et al. (2004), we found significant evaluation differences for four of the five subscales across parent and teacher SDQ scores. On average, parents score their children significantly lower with regard to emotional symptoms and peer problems, but evaluate the prosocial behavior of their children significantly better than teachers. In contrast, parents perceive their children to be more affected by conduct problems than teachers. However, without strict invariance for the SDQ across parents and teachers, it cannot be assumed that these differences reflect evaluation differences between parents and teachers.

### Data Analysis

The statistical analysis consists of two parts. First, to analyze the internal structure of the parent and teacher
SDQ, several confirmatory factor analyses (CFAs) were conducted. Second, using multigroup confirmatory factor analysis (MGCFA), we tested for measurement invariance across parent and teacher SDQ versions.

Various expectations about the factor structure of the SDQ are discussed in the literature. While Ruchkin et al. (2007, 2008), Zhang et al. (2009), Burstein and Schmitz (2012), Svedin and Priebe (2008), and Van Roy et al. (2008) have shown that the five-factor solution fits the best, other studies have found the three-factor solution to be the most appropriate (Dickey & Blumberg, 2004; Di Riso et al., 2010; Percy et al., 2008). For the three-factor solution, emotional problems and peer problems are referred to as the scale internalizing symptoms, while conduct problems and hyperactivity are referred to as a scale named externalized problems (Ortuño-Sierra et al., 2015).

Regarding the factor structure, Van de Looij-Jansen et al. (2011), Chiorri et al. (2015), and Ortuño-Sierra et al. (2015) pointed out that the use of reverse-worded items could be difficult as these items may more precisely measure a diverse construct. Because of lower correlations with the remaining items of the problem scales, the reverse-worded items might produce problems with the factor structure. To overcome this, Van de Looij-Jansen et al. (2011) and Ortuño-Sierra et al. (2015) proposed allowing cross-loadings of the reverse-worded items to the prosocial behavior factor. Moreover, they argued that the use of only five items to cover the extensive area of psychological problems might be problematic as items may represent connected but diverse subareas, which lead to less consistent scales. As a consequence, they recommended allowing the correlation between SDQ items with similar content. The allowing of cross-loadings of the reverse-worded items, as well as allowing the correlation between SDQ items with similar content, resulted in a significantly better model fit with the SDQ scores in both studies.

Therefore, in a first step using confirmatory factor analysis (CFA) we tested for the following eight solutions:

1. The original five-factor model,
2. The five-factor model with reverse-worded items allowed to cross-load on the prosocial factor,
3. The five-factor model with correlated errors allowing for items with similar content,
4. The five-factor model with correlated errors and the reverse-worded items allowed to cross-load on the prosocial factor,
5. The three-factor model,
6. The three-factor model with reverse-worded items allowed to cross-load on the prosocial factor,
7. The three-factor model with correlated errors allowing for items with similar content, and
8. The three-factor model with correlated errors and the reverse-worded items allowed to cross-load on the prosocial factor.

Due to the categorical data structure of the SDQ, the scores are not normal distributed (Table 2). Therefore, we used the robust mean-adjusted weighted least square method (WLSMV) for the estimation of all the following statistical analyses. All eight models were run for the whole sample, as well as for parent and teacher SDQ scores alone.

In a second step, we tested for measurement invariance between parent and teacher SDQ scores for the CFA model with the best fit to the data. Using multigroup confirmatory factor analysis (MGCFA), we set cross-group constraints on parameters and compared model fits from less restricted models with more restricted models (Byrne & Stewart, 2006; Cieciuch et al., 2015; Davidov, 2010; Steenkamp & Baumgartner, 1998). To test for configural invariance we ran a model in which all factor loadings and thresholds were estimated for all groups simultaneously without any equality constraints (Davidov, 2010; Horn & McArdle, 1992).

"If a fixed parameter for an outcome has a modification index or a derivative of zero, it will not be identified if it is free" (Muthén & Muthén, 2007, p. 470). We therefore had to restrict the lower threshold of items 7 and 9, as well as the upper thresholds of items 1-6, 8, and 10-25 to be equal between groups.

To evaluate the existence of measurement invariance we used the comparative fit index (CFI), as well as the Tucker-Lewis index (TLI), the root mean square residual

### Table 2. Mean and standard deviation of SDQ subscales and total difficulties score

<table>
<thead>
<tr>
<th>Subscale</th>
<th>Parent mean (SD)</th>
<th>Teacher mean (SD)</th>
<th>Mean difference</th>
<th>Total mean (SD)</th>
<th>Total skewness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotional symptoms</td>
<td>1.54 (1.64)</td>
<td>1.71 (2.04)</td>
<td>-0.167</td>
<td>1.62 (1.85)</td>
<td>1.66</td>
</tr>
<tr>
<td>Conduct problems</td>
<td>1.71 (1.54)</td>
<td>1.44 (1.76)</td>
<td>0.275***</td>
<td>1.58 (1.66)</td>
<td>1.30</td>
</tr>
<tr>
<td>Hyperactivity</td>
<td>3.61 (2.39)</td>
<td>3.43 (2.73)</td>
<td>0.175**</td>
<td>3.52 (2.57)</td>
<td>0.60</td>
</tr>
<tr>
<td>Peer problems</td>
<td>1.28 (1.48)</td>
<td>1.64 (1.93)</td>
<td>-0.360**</td>
<td>1.46 (1.73)</td>
<td>1.44</td>
</tr>
<tr>
<td>Prosocial</td>
<td>7.63 (1.76)</td>
<td>6.75 (2.42)</td>
<td>0.879***</td>
<td>7.19 (2.16)</td>
<td>-0.66</td>
</tr>
<tr>
<td>Total difficulties</td>
<td>8.15 (4.87)</td>
<td>8.23 (5.87)</td>
<td>-0.077</td>
<td>8.19 (5.39)</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Notes. N<sub>parents</sub> = 1,057; N<sub>teachers</sub> = 1,057; Wilcoxon signed-rank test. ***p < 0.001; **p < 0.01.
(RMSEA), and the weighted root mean square residual (WRMR). To assume that configural invariance is given, the CFI and TLI values must be at least greater than 0.90 and the RMSEA should be smaller than 0.08 (Acock, 2013). Once configural invariance is given, the model fit is used as the baseline against which the more restricted model is compared.

To evaluate the existence of scalar and strict measurement invariance, we used the criteria suggested by Chen (2007). Using simulation studies, Chen (2007) showed that a change of −0.01 compared to the configural invariance model in the CFI value, together with a simultaneous increase of the RMSEA larger than 0.015, points to a lack of invariance. Due to the ordered scale level of the SDQ items, factor loadings, as well as thresholds, influence the probabilities of the factor indicators. Therefore, to test for metric and scalar invariance we restricted both parameters simultaneously in the next step (Ortuño-Sierra et al., 2015; Stone et al., 2013). To test for strict measurement invariance, starting from the scalar measurement invariance model in the previous step, the residual variances were constrained to be equal across groups.

All analyses were run with Mplus 5 (Muthén & Muthén, 2007).

Results

Internal Structure of the SDQ Scores

The results of the confirmatory factor analyses for the eight hypothetical expectations are displayed in Table 3.

The goodness-of-fit statistic indicates a poor fit for the original five-factor and baseline three-factor models. In contrast, the five-factor model and three-factor model with reverse-worded items allowed to cross-load on the prosocial factor lead to an increase in the CFI value to between 0.044 and 0.051. However, a moderate fit is only displayed for the five-factor solution with reverse-worded items allowed to cross-load on the prosocial factor (CFI = 0.914). As some correlation errors were found in the original five-factor model, we allowed the correlation between items with similar content (2 & 10, 6 & 24, 21 & 1, 20 & 9, 23 & 6, 11 & 6, 24 & 6) for the five-factor and three-factor models. This modification (models 3 and 7) leads to a significant increase in the model fit in both solutions. However, as before, only the five-factor model obtains a moderate model fit (CFA = 0.914) with this modification.

The greatest increase in model fit for the five-factor and three-factor solutions (ΔCFI = 0.068 for the five-factor solution and ΔCFI = 0.083 for the three-factor solution) is seen when both modifications (models 4 and 8) are combined. While both factor solutions reach a moderate model fit for the whole sample and the separate parent and teacher scores after combining modifications, the model fit for the five-factor solution is still better (ΔCFI ≥ 0.028) than for the three-factor solution. With CFI values between 0.938 and 0.961, TLI values higher than 0.95 and RMSEA values smaller than 0.08 for the whole sample, as well as for teachers and parents alone, model 4 shows a better fit to the data than all other tested hypothetical dimensions. Therefore, the following testing for measurement invariance has been based on model 4 (Table 3).

Measurement Invariance of the SDQ Scores Across Teachers and Parents

To test for configural invariance, we first fitted a multigroup confirmatory factor analysis that imposes the equivalent form on all the relations without imposing any equality restrictions (Acock, 2013). The goodness-of-fit indices (model 1, Table 4) with CFI and TLI values larger than 0.95 and a RMSEA smaller than 0.06 indicate that the same sets of items are associated with the same factors (subscals) for parents and teachers (Van de Looij-Jansen et al., 2011).

Next, we tested for scalar invariance. We restricted thresholds and factor loadings to be equal across parent and teacher scores. The goodness-of-fit statistics between the configural invariance model (model 1) and the scalar invariance model (model 2) indicate that scalar invariance is given. All goodness-of-fit indices demonstrate a worse but not significant fit for the scalar invariance model (ΔCFI = −0.006; ΔTLI = −0.003; ΔRMSEA = 0.002) when compared to the configural invariance model. Hence, we can assume that the five latent variables have the same meaning for teachers and parents (Acock, 2013). As a result, the latent mean differences between parents and teachers (Table 5) indicate the existence of real group evaluation differences between parents and teachers.

Finally, we constrained the residual variance to be equal across parent and teacher SDQ scores (model 3). Even though this model reveals a poorer fit compared to model 1 (ΔCFI = −0.007; ΔRMSEA = 0.003), the drop is not significant (Table 4). Therefore, the results imply that a meaningful comparison across observed parent and teacher SDQ scores is tenable.

Besides testing for measurement invariance and the evaluation of the overall model fit, we were also interested in assessing the model based on item and construct level. We therefore checked for significant and required item loadings (item level) and calculated factor reliability scores (construct level). The standardized factor loadings and the factor reliability scores for the strict measurement
invariance model for parents and teachers are displayed in Table 6. Table 6 indicates that all factor loadings are statistically significant at the 0.001 level. Additionally, all loadings are restricted in the expected direction. With scores ranging between 0.62 (peer problems parent scores) and 0.84 (prosocial teacher scores), all factors reach an acceptable reliability. However, the internal consistency seems to be higher for the teacher version than for the parent one, which was also indicated in Table 1 by Cronbach’s α. Nevertheless, some loadings of the reverse-worded items do not reach the required loadings from 0.3 to the prosocial behavior factor (Knafl & Grey, 2007). Aside from that, all loadings range between 0.327 and 0.93.

Table 3. Goodness-of-fit indices CFA

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Original five-factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>2,076.187</td>
<td>125</td>
<td>0.870</td>
<td>0.929</td>
<td>0.086</td>
</tr>
<tr>
<td>Only parents</td>
<td>710.705</td>
<td>132</td>
<td>0.891</td>
<td>0.928</td>
<td>0.064</td>
</tr>
<tr>
<td>Only teachers</td>
<td>1,022.945</td>
<td>90</td>
<td>0.906</td>
<td>0.946</td>
<td>0.099</td>
</tr>
<tr>
<td>2. Five-factor with reverse-worded items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>15,078.943</td>
<td>68</td>
<td>0.914</td>
<td>0.954</td>
<td>0.070</td>
</tr>
<tr>
<td>Only parents</td>
<td>466.299</td>
<td>130</td>
<td>0.937</td>
<td>0.958</td>
<td>0.049</td>
</tr>
<tr>
<td>Only teachers</td>
<td>723.457</td>
<td>88</td>
<td>0.936</td>
<td>0.962</td>
<td>0.083</td>
</tr>
<tr>
<td>3. Five-factor with CE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>1,417.799</td>
<td>126</td>
<td>0.914</td>
<td>0.954</td>
<td>0.070</td>
</tr>
<tr>
<td>Only parents</td>
<td>484.578</td>
<td>131</td>
<td>0.933</td>
<td>0.956</td>
<td>0.051</td>
</tr>
<tr>
<td>Only teachers</td>
<td>746.577</td>
<td>90</td>
<td>0.934</td>
<td>0.962</td>
<td>0.083</td>
</tr>
<tr>
<td>4. Five-factor with CE and reverse-worded items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>1,062.831</td>
<td>126</td>
<td>0.938</td>
<td>0.966</td>
<td>0.059</td>
</tr>
<tr>
<td>Only parents</td>
<td>334.267</td>
<td>130</td>
<td>0.961</td>
<td>0.974</td>
<td>0.039</td>
</tr>
<tr>
<td>Only teachers</td>
<td>591.512</td>
<td>88</td>
<td>0.949</td>
<td>0.970</td>
<td>0.074</td>
</tr>
<tr>
<td>5. Baseline three-factor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>2,859.054</td>
<td>111</td>
<td>0.817</td>
<td>0.888</td>
<td>0.108</td>
</tr>
<tr>
<td>Only parents</td>
<td>5,383.290</td>
<td>128</td>
<td>0.833</td>
<td>0.887</td>
<td>0.081</td>
</tr>
<tr>
<td>Only teachers</td>
<td>1,488.325</td>
<td>77</td>
<td>0.858</td>
<td>0.904</td>
<td>0.132</td>
</tr>
<tr>
<td>6. Three-factor with reverse-worded items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>2,093.945</td>
<td>114</td>
<td>0.868</td>
<td>0.921</td>
<td>0.091</td>
</tr>
<tr>
<td>Only parents</td>
<td>742.535</td>
<td>126</td>
<td>0.884</td>
<td>0.920</td>
<td>0.068</td>
</tr>
<tr>
<td>Only teachers</td>
<td>1,134.299</td>
<td>81</td>
<td>0.894</td>
<td>0.932</td>
<td>0.111</td>
</tr>
<tr>
<td>7. Three-factor with CE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>15,078.943</td>
<td>112</td>
<td>0.868</td>
<td>0.920</td>
<td>0.092</td>
</tr>
<tr>
<td>Only parents</td>
<td>706.705</td>
<td>128</td>
<td>0.891</td>
<td>0.926</td>
<td>0.065</td>
</tr>
<tr>
<td>Only teachers</td>
<td>1,186.910</td>
<td>79</td>
<td>0.888</td>
<td>0.926</td>
<td>0.115</td>
</tr>
<tr>
<td>8. Three-factor with CE and reverse-worded items</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Whole sample</td>
<td>1,459.566</td>
<td>115</td>
<td>0.910</td>
<td>0.947</td>
<td>0.074</td>
</tr>
<tr>
<td>Only parents</td>
<td>513.122</td>
<td>127</td>
<td>0.927</td>
<td>0.950</td>
<td>0.054</td>
</tr>
<tr>
<td>Only teachers</td>
<td>819.137</td>
<td>81</td>
<td>0.926</td>
<td>0.952</td>
<td>0.093</td>
</tr>
</tbody>
</table>

Notes. $N_{\text{parents}} = 1,057$; $N_{\text{teachers}} = 1,057$. CE = correlation errors; df = WLSMV does not calculate model degrees of freedom in the standard way. Instead, degrees of freedom are estimated to approximate a $\chi^2$ distribution (Muthén & Muthén, 2007).

Table 4. Goodness-of-fit indices for measurement invariance of the SDQ across teachers and parents

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$</th>
<th>df</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>$\Delta$CFI</th>
<th>Model comparison</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural invariance (1)</td>
<td>957.228</td>
<td>216</td>
<td>0.955</td>
<td>0.973</td>
<td>0.057</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scalar invariance (2)</td>
<td>1,066.939</td>
<td>226</td>
<td>0.949</td>
<td>0.970</td>
<td>0.059</td>
<td>$-0.006$</td>
<td>2 versus 1</td>
</tr>
<tr>
<td>Strict invariance (3)</td>
<td>1,074.968</td>
<td>226</td>
<td>0.948</td>
<td>0.970</td>
<td>0.060</td>
<td>$-0.007$</td>
<td>3 versus 1</td>
</tr>
</tbody>
</table>

Note. $N_{\text{parents}} = 1,057$; $N_{\text{teachers}} = 1,057$. 

This document is copyrighted by the American Psychological Association or one of its allied publishers. This article is intended solely for the personal use of the individual user and is not to be disseminated broadly.
## Discussion and Conclusion

The main purpose of this study was to examine the internal structure of the parent and teacher SDQ in a sample of German preschoolers and to test whether the SDQ is measurement invariant across parents and teachers.

Findings regarding the internal structure indicate that the original five-factor structure is not convincing for the parent and teacher SDQ. For this reason the use of the original factor structure is questionable. A good model fit could only be reached after allowing for the correlation between SDQ items with similar content and after allowing for cross-loadings of the reverse-worded items to the prosocial behavior factor. This finding is in line with those of previous studies of Van de Looij-Jansen et al. (2011) and Ortuño-Sierra et al. (2015) for the SDQ self-reported version. Both conclude that the better fit of models with cross-loadings indicates that responses to the reverse-worded items reflect both the presence of a specific problem and a positive response product. Thus, it is impossible to conclude whether a problem behavior exists or if the outcome is just the result of an informant response style to positive wording. Due to these results, the question arises whether replacing the reverse-worded SDQ items with corresponding problem items or the obligatory use of the extended prosocial behavior scale would be beneficial.

Our findings regarding the question of measurement invariance across parents and teachers suggest that strict measurement invariance across both groups is tenable. Hence, the displayed significant differences in SDQ scores of parent and teacher ratings (Table 2) are readily comparable and interpretable. This is an important finding since psychological problems are supposed to be highly context specific (Achenbach, 1991; Goodman, Ford, et al., 2000; Stone et al., 2010). Furthermore, the evidence of strict measurement invariance allows only the systematic analysis of the factors that influence mean differences between parent and teacher ratings.

### Differences in parent and teacher ratings might be explained by (1) differences between informants, (2) disparities in the rating context, (3) the nature of the observed behavior, and (4) the age of the observed child (De Los Reyes & Kazdin, 2005).

Regarding the first point, it might be that parents and teachers have dissimilar understandings of problem behavior. In addition, parents may be more vulnerable to overstating a positive behavior and underreporting problem behaviors (Kuschel, Heinrichs, Betram, Naumann,
& Hahlweg, 2007; Rademacher, Koglin, & Petermann, 2016). In contrast, teachers probably rate pupils relative to other students in the class (Kristoffersen & Smith, 2013). The special demands that kindergarten places on children might be an example for the influence of environmental context on the rating. For example, because a child has to share the attention of the caregiver with other children, a higher degree of self-regulatory skills is required in kindergarten. With regard to the third point, Grietens et al. (2004) pointed out that the probability of rater agreement is higher for observed externalized than for observed internalized behavior. Regarding the last explanation, Achenbach, McConaughy, and Howell (1987) have shown that agreement between informant ratings is greater for younger children (6–11 years) than for adolescents (12–19 years). The authors interpreted this result as suggesting that younger children’s behavior may itself be more cross-situationally consistent (De Los Reyes & Kazdin, 2005).

It is acknowledged that this study has a few limitations. Firstly, measurement invariance across parent and teacher SDQ scores can only be assumed with regard to the evaluation of preschoolers aged 3–6 years. This limitation is important because Achenbach et al. (1987) have shown that agreement between parents and teachers is higher for young children. Secondly, since Ortuño-Sierra et al. (2015) found that the self-reported version of the SDQ reaches only partial measurement invariance across countries, it cannot be assumed that the parent and teacher SDQ are measurement invariant across countries. Consequently, we cannot conclude that the SDQ scores of parent and teacher ratings for preschoolers are comparable in other countries. Finally, since the analysis is not based on a representative sample of German preschoolers we cannot assume that all SDQ items have the same meaning for all parents and teachers in Germany. Lastly, based on this study it is not possible to determine which of the four explanations caused the rating differences between parents and teachers.

Besides these shortcomings, the strength of this study is that parents and teachers evaluated the same children, leading to the conclusion that evaluation differences are not the result of different samples, but rather the result of rating differences between parents and teachers. Therefore, a content interpretation of evaluation differences between parent and teacher SDQ scores is permitted and rating differences between parents and teachers can be used for diagnostic purposes.

References


Acoc, A. (2013). Discovering structural equation modeling using Stata. College Station, TX: Stata Press.


Appendix

Items Parent and Teacher SDQ Questionnaires

1. “Considerate of other people’s feelings”
2. “Restless, overactive, cannot stay still for long”
3. “Often complains of headaches, stomach aches or sickness”
4. “Shares readily with other children (treats, toys, pencils, etc.)”
5. “Often has temper tantrums or hot tempers”
6. “Rather solitary, tends to play alone”
7. “Generally obedient, usually does what adults request”
8. “Many worries, often seems worried”
9. “Helpful if someone is hurt, upset or feeling ill”
10. “Constantly fidgeting or squirming”
11. “Has at least one good friend”
12. “Often fights with other children or bullies them”
13. “Often unhappy, down-hearted or tearful”
14. “Generally liked by other children”
15. “Easily distracted, concentration wanders”
16. “Nervous or clingy in new situations, easily loses confidence”
17. “Kind to younger children”
18. “Often argumentative with adults”
19. “Picked on or bullied by other children”
20. “Often volunteers to help others (parents, teachers, other children)”
21. “Can stop and think things out before acting”
22. “Can be spiteful to others”
23. “Gets on better with adults than with other children”
24. “Many fears, easily scared”
25. “Sees tasks through to the end, good attention span”


Received May 12, 2016
Revision received October 4, 2016
Accepted October 9, 2016
Published online November 29, 2017

Jana Rogge
Department of Special Education and Rehabilitation
Carl v. Ossietzky University Oldenburg
26111 Oldenburg
Germany
jana.rogge@uni-oldenburg.de