Article

Measurement Equivalence and Convergent Validity of a Mental Health Rating Scale

Assessment
2020, Vol. 27(8) 1901–1913

DOI: 10.1177/1073191118803159
journals.sagepub.com/home/asm
SAGE

Sanne C. Smid[1*], Joop J. Hox[1*], Einar R. Heiervang[2,3], Kjell Morten Stormark[4,5], Mari Hysing[4], and Tormod Bøe[4]

Abstract

Emotional and behavioral problems among children and adolescents may be studied using the Strengths and Difficulties Questionnaire, containing five subscales, based on ratings by parents, teachers, or adolescents themselves. We investigate two measurement issues using data from a longitudinal sample of 8,806 participants aged 7 to 9 years and 11 to 13 years from the Bergen Child Study in Bergen, Norway. First, convergent validity of parent and teacher ratings is studied using a multitrait–multimethod approach. Second, longitudinal measurement equivalence is studied using confirmatory factor analysis, which requires us to deal with the considerable attrition. The multitrait–multimethod indicates not only good convergent validity but also considerable method variance for parents and teachers. The reliability and validity of some subscales are relatively low. Attrition analysis indicates that attrition is not missing completely at random, but estimation assuming missing at random makes no real difference. We conclude that assuming missing completely at random is acceptable. Comparing ratings by parents and teachers results in partial scalar equivalence. In addition, all subscales exhibit (partial) longitudinal scalar measurement equivalence. We recommend using latent variable modeling and not summated scales for longitudinal modeling using the Strengths and Difficulties Questionnaire.

Keywords

measurement equivalence, measurement invariance, convergent validity, MTMM, Mental Health Rating Scale, SDQ

Emotional and behavioral problems are prevalent among children (Costello, Foley, & Angold, 2006), and a majority of those experiencing mental health problems during their lifetime debut these during childhood and adolescence (Kessler et al., 2005). To facilitate early identification and prevention, identifying developmental trajectories of emotional and behavioral problems is crucial. Longitudinal studies of mental health in childhood are well suited for these purposes, because they allow identifying patterns of continuity and change across development, as well as antecedents of risk factors and outcomes.

The Strengths and Difficulties Questionnaire (SDQ) is a screening instrument designed for monitoring of emotional and behavioral problems in children and adolescents aged 2 to 17 years (Goodman, Ford, Simmons, Gatward & Meltzer, 2000). The SDQ consists of five scales, each indicated by five items. The instrument construction was based on a theoretical framework and consists of four subscales indicating problems: hyperactivity/inattention, conduct problems, emotional symptoms, and peer problems, as well as one prosocial subscale (Goodman, 1997, 1999). The SDQ is a multi-informant instrument, with versions for teachers and parents as raters, and a self-report version intended for adolescents aged 11 to 16 years. The SDQ has been translated into more than 80 languages, and is an attractive instrument because it is publicly available at no cost (www.sdqinfo.org).

The five-factor model of the SDQ has received considerable support (for a review, see Stone, Otten, Engels, Vermulst, & Janssens, 2010), although sometimes minor method factors are also found (van de Looij-Jansen, Goedhart, de Wilde, & Treffers, 2011). Previous investigations from the same population as the current sample have demonstrated that the five-factor solution fitted both parent and teacher ratings of 7- to 9-year-olds (Sanne, Torsheim, Heiervang, Stormark, & Morten, 2009) and also self-ratings of 16- to 18-year-old girls and boys (Bøe, Hysing, Skogen, & Breivik, 2016).

[1]Utrecht University, Utrecht, the Netherlands
[2]University of Oslo, Oslo, Norway
[3]Oslo University Hospital, Oslo, Norway
[4]Uni Research Health, Bergen, Norway
[5]University of Bergen, Bergen, Norway
*Shared first authorship

Corresponding Author:
Sanne C. Smid, Department of Methodology and Statistics, Utrecht University, P.O. Box 80.140, 3805 TC Utrecht, the Netherlands.
Email: s.c.smid@uu.nl

Although the SDQ has been used in longitudinal studies (e.g., Becker, Rothenberger, Sohn, & BELLA Study Group, 2015; Sayal, Heron, Golding, & Emond, 2007), the construct validity has primarily been investigated in cross-sectional samples assessing the construct validity of the SDQ across informants, age, and gender (cf. Bøe et al., 2016; Chiorri, Hall, Casely-Hayford, & Malmberg, 2016; He, Burstein, Schmitz, & Merikangas, 2013, Palmieri & Smith, 2007; Rønning, Handegaard, Sourander, & Mørch, 2004; Sanne et al., 2009; van de Looij-Jansen et al., 2011). However, in a study of rapidly developing children, it is also important to assure that we are in fact measuring the same construct over time. Few studies have looked at measurement invariance of SDQ longitudinally. In a sample of 180 German children, measurement invariance over time for the SDQ teacher version was demonstrated (DeVries, Gebhardt, & Voß, 2017). Similarly, strong longitudinal factorial invariance was found for the five-factor model for parent report in preschool children (Sosu & Schmidt, 2017).

An additional (longstanding) issue is whether different raters (parents, teachers) are actually indicating the same construct (Achenbach, McConaughy, & Howell, 1987; Munkvold, Lundervold, Lie, & Manger, 2009). Correlations between parent and teacher SDQ ratings were found to show only moderate agreement in a multinational sample of 6- to 11-year-old children from seven European countries (Cheng et al., 2015). However, in a study of first graders at risk of educational failure, using a multitrait–multimethod (MTMM) approach as well as confirmatory factor analyses, Hill and Hughes (2007) found evidence for convergence among parent, teacher, and peer ratings; parents ratings contained relatively less trait variance and displayed the strongest method effects.

For clinical practice and in research the use of both parent and teacher report are needed when screening for mental health problems, and when studying development of mental health in longitudinal samples, it is critical to assure that what we observe are meaningful changes in behavior and not artifacts of measurement. There are few studies with sophisticated analyses of parent–teacher convergence using large population samples of children, and very few have investigated longitudinal measurement invariance of the SDQ. Using longitudinal data from the first two waves of the Bergen Child study, a prospective cohort study of children's mental health from primary school age to adolescence, we address two measurement related issues for the SDQ. The first issue is the convergent validity of parent and teacher ratings; do they measure the same construct? The second issue is whether there is longitudinal measurement equivalence. In order to investigate the longitudinal measurement equivalence, we need to deal with a third issue, which is the substantial attrition in our data.

The next section describes the data at our disposal. We then (a) present analysis models for convergent and discriminant validity for parent and teacher data, (b) analyze the attrition in our data, and (c) assess longitudinal measurement equivalence taking into account the findings about the attrition process.

Method

Data

In this study, we employed data from the Bergen Child Study, a longitudinal population-based study of mental health launched in 2002. The main aim of the study was to produce prevalence data for mental health problems, including comorbidity, risk, and protective factors, as well as the use of health and educational services. Three age cohorts in the municipality of Bergen are included in four consecutive waves of data collection. The present study is based on the two first waves of the study at ages 7 to 9 (2002) and 11 to 13 years (2006). Parents and teachers were informants in the two first waves, and in the second wave, the children's self-report was also included. In the present study, the data are restricted to parent and teacher reports from the first two waves.

In a previous investigation to determine the representativeness of the study population using the same sample as the current, the magnitude of nonresponse bias was estimated by comparing teacher ratings of participating and nonparticipating children (Stormark, Heiervang, Heimann, Lundervold, & Gillberg, 2008). While there were more emotional and behavioral problems in nonparticipating children, the differences were small in magnitude, suggesting that the current sample is representative of the total population (Stormark et al., 2008).

In Wave 1, 7,183 children participated in the study, in Wave 2, 5,647 children participated. Table 1 specifies the number of children and available ratings in more detail.

Instrument

The SDQ (Goodman, 1997, 1999) is a multi-informant screening questionnaire of emotional and behavioral problems for children between 4 and 16 years. It consists of 25 items describing positive and negative attributes of children that can be allocated to five subscales with five items each: (a) emotional symptoms (EMOT), (b) conduct problems (COND), (c) hyperactivity–inattention problems (ADHD), (d) peer relationship problems (PEER), and (e) prosocial behavior (PROS). A total difficulty score is computed by combining the first four subscales. Each subscale is scored on a three-point scale; "not true," "somewhat true," and "certainly true" with total subscales scores each ranging from 0 to 10 and a total difficulties score from 0 to 40.

Table 1. Number of Participants in Waves 1 and 2, and Specification of the Rater Who Supplied the Data.

Waves	*n*	Only parent ratings	Only teacher ratings	Both parent and teacher ratings
Wave 1	7,183	398 (5.5%)	309 (4.3%)	6,476 (90.2%)
Wave 2	5,647	82 (1.5%)	461 (8.2%)	5,104 (90.4%)

Note. The total number of participants is 8,806. Some participants provided data at the first wave and failed to provide data at the second wave. For other participants, it is the other way around: data are provided at the second wave, but not at the first wave. As a consequence, the number of cases reported in different analyses do not always match up.

Ethics

The study was approved by the Regional Committee for Medical and Health Research Ethics in Western Norway. Informed consent was obtained from all parents included in this study.

Analysis Models

The SDQ items have three response categories: "not true," "somewhat true," and "certainly true." The response distribution is typically skewed, with "not true" chosen the most and "certainly true" the least, the exception being five positively phrased items and prosocial scale items, where scoring is reversed. The estimation method needs to take the ordinal categorical nature of the data into account. Based on the attrition analysis (reported in the results section), we use a robust weighted least squares estimator (WLSM; L. K. Muthén & B. O. Muthén, 1998-2012). The ordinal categorical nature of the data has a consequence for the measurement equivalence testing. When testing for measurement equivalence with continuous data, there is a separate step for *metric invariance*, estimating a model with only the factor loadings constrained equal across waves. This is also possible with ordered categorical data (Millsap, 2011; Millsap & Yun-Tein, 2004), but it does not add information, because with categorical indicators a change in the thresholds implies a change in the loadings. Following recommendations by B. O. Muthén (2013), we do not include a separate step for *metric equivalence* in our analyses on the ordinal categorical data.

Convergent and Discriminant Validity

The convergent validity of the parent and teacher ratings is analyzed in Wave 1 using an MTMM approach (Campbell & Fiske, 1959). Marsh and Grayson (1995, p. 177) describe traits as "attributes such as multiple abilities, attitudes, behaviors, or personality characteristics," whereas methods "refer broadly to multiple test forms, methods of assessment, raters, or occasions." In our case, the traits are the five subscales included in the SDQ, which are modeled as continuous variables, and the methods are the raters, the parents, and teachers. The analysis model is a confirmatory factor model, first described by Jöreskog (1971). Marsh and Grayson (1995) describe this model as the correlated traits–uncorrelated methods model. Figure 1 presents the path diagram for the correlated traits–uncorrelated methods (CTUM) model for our data.

The path diagram in Figure 1 is described by the following confirmatory factor analysis equation:

$$Y_{ij} = I_i + t_{1i}\theta_{1j} + t_{2i}\theta_{2j} + \ldots + t_{5i}\theta_{5j} + m_{1i}\mu_{1j} + m_{2i}\mu_{2j} + e_{ij} \quad (1)$$

where Y_{ij} is the score of subject j on subscale i, I_i is the intercept of subscale i, $t_{1i} \ldots t_{5i}$ are the loadings of subscale i on trait $1 \ldots 5$, $\theta_{1j} \ldots \theta_{5j}$ are the scores of subject j on trait $1 \ldots 5$, m_{1i} and m_{2i} are the loadings of subscale i on method 1 and 2, μ_{1j} and μ_{2j} are the score of subject j on method 1 and 2, and e_{ij} is the residual error of subject j for subscale i. The factor loading of indicators who do not belong to a specific trait or method are constrained to 0. The intercepts I_i are implicit in Figure 1, and they are necessary because incomplete data are analyzed using full information maximum likelihood (Enders, 2010). The scale of the factors is defined by constraining their means to 0 and the variances to 1, and all intercepts and all unconstrained loadings are freely estimated.

Convergent validity refers to the overlap of alternative methods intended to measure the same construct, but having different sources of systematic error (Campbell & Fiske, 1959; Marsh & Grayson, 1995; Widaman, 1985). Large trait factor loadings are an indication of convergent validity (Marsh & Grayson, 1995).

There is discriminant validity when the subscales are indeed measuring different constructs, which implies that the latent constructs are not too highly correlated (Campbell & Fiske, 1959; Marsh & Grayson, 1995; Widaman, 1985). Marsh and Grayson (1995) indicate that large trait correlations, especially correlations near 1.0, signify a lack of discriminant validity. T. D. Little (2013) prefers that for MTMM models the trait correlations should be below .7, but states that in practice in MTMM models trait correlations around .8 are not unusual. In addition, the MTMM allows assessing method effects. We have a method effect if the observed correlations among different traits measured, with the same

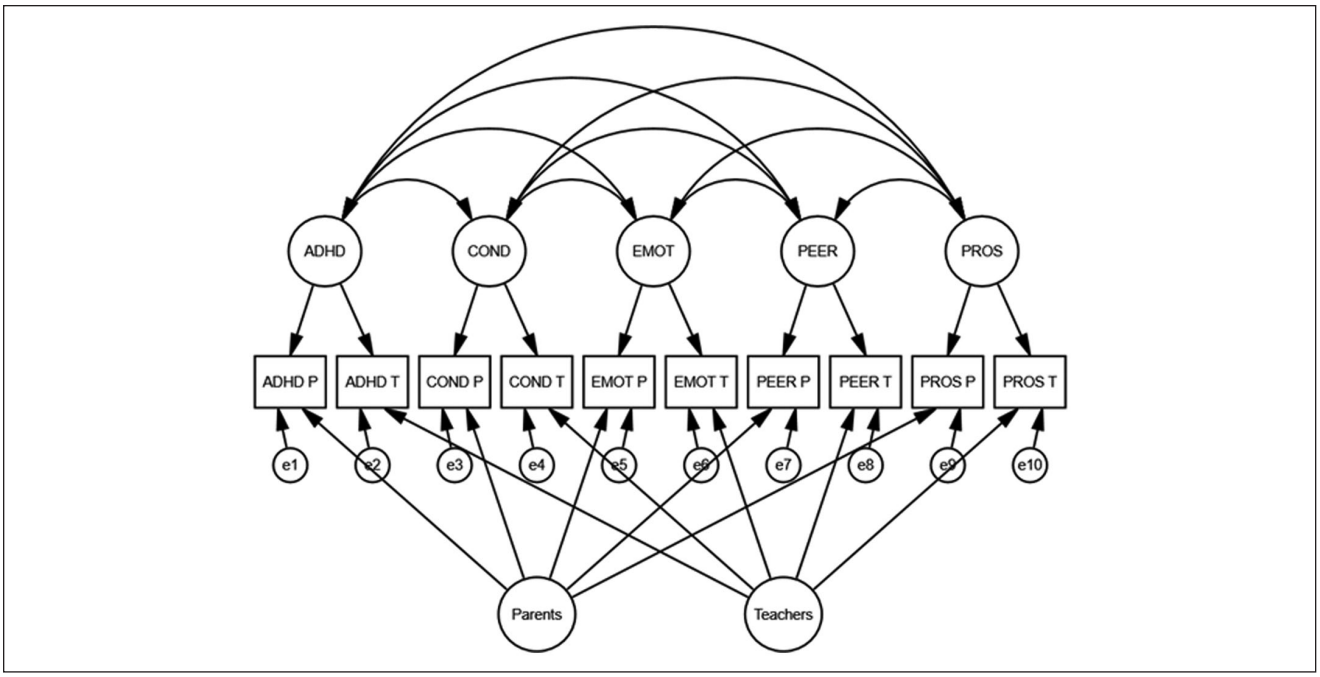

Figure 1. The CTUM MTMM model for the SDQ data.
Note. CTUM = correlated traits–uncorrelated methods; MTMM = multitrait–multimethod; SDQ = Strengths and Difficulties Questionnaire; P = parent rating, T = teacher rating; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.

method, are increased by the method. Thus, large method factor loadings suggest a method effect (Campbell & Fiske, 1959; Marsh & Grayson, 1995; Widaman, 1985).

Attrition

In panel data, attrition or panel dropout is generally a problem. In the missing data literature, three missing data mechanisms are distinguished: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR; R. J. A. Little & Rubin, 2002; Rubin, 1976). These mechanisms specify the relationship between the observed values in the data set and the missing values. Missing values are MCAR when the probability P of a value to be missing is totally unrelated to all observed or unobserved variables in the data set. In other words, if we define a response indicator R where 0 represents that an outcome variable Y is missing and 1 represents that Y is observed, we have MCAR if R is not related to the observed variables and the (unknown) values of the missing data (Schafer & Graham, 2002).

The second mechanism is MAR. When the data are MAR, there is a relation between the response indicator R and the observed data, but no relation between the response indicator R and the (unknown) missing values. When the data are MAR, estimation methods are available such as full information maximum likelihood that will produce unbiased estimates.

The final mechanism is MNAR. We have MNAR, when there is a relation between the probability of a value to be missing and the (unknown) missing value itself. It is not possible to identify whether the data are MAR or MNAR, because the difference is whether the missingness depends on the unobserved value, and that information is by definition not available.

For an elaboration of these mechanisms and appropriate analysis methods, we refer to Rubin (1976), Schafer and Graham (2002), and van Buuren (2012). The mechanisms described above refer to both unit nonresponse (entire observational unit fails to provide data) and item nonresponse (values are missing for some of the variables). However, in our data, the major missing data pattern is attrition, followed by cases that fail to provide data at the first wave, but are captured at the second wave. As a consequence, the missing data analysis focuses on these two patterns. In addition, due to the missing data at different occasions, the number of cases reported in different analyses does not always match up. A detailed breakdown of the number of participants at each wave can be found in our analysis report (https://bit.ly/2LhiRCA).

Parent–Teacher and Longitudinal Measurement Equivalence

The SDQ uses different sources of information: parents and teachers rate the children, and older adolescents may rate

themselves. The main focus in the current article is on *longitudinal* measurement equivalence. Measurement equivalence between parents and teachers is analyzed and briefly reported for Wave 1 and Wave 2 separately, because these waves include the most respondents.

Measurement equivalence or measurement invariance is usually considered when different groups are compared. It is generally investigated using multigroup confirmatory factor analysis. In longitudinal research, measurement equivalence is usually less of an issue, because we are measuring the same subjects using the same instruments. However, measurement equivalence should not be assumed but investigated, especially when the goal is to study development.

Measurement equivalence over time follows the same steps as measurement equivalence across groups, although the analysis details differ. The strongest form of measurement equivalence is scalar equivalence or strong measurement invariance. Mellenbergh (1989) defines this as implying that the relationship between the observed score and the unobserved latent score of a subject does not depend on group membership. Translated to longitudinal designs, it means that the relationship between the observed score and the unobserved latent score of a subject does not depend on the measurement occasion. Scalar equivalence or strong measurement invariance implies that both factor loadings and intercepts can be constrained equal across time. A weaker form of measurement equivalence is metric equivalence or weak measurement invariance. This implies that loadings can be constrained equal across time but intercepts cannot. Finally, the weakest form of equivalence is configural equivalence, which implies that the same factor structure holds across time, but neither loadings nor intercepts can be constrained equal.

For a valid comparison of observed (summated) scores, we need full scalar equivalence. For a valid comparison of latent factor means, it is sufficient to have partial scalar equivalence. The minimal requirements for partial invariance are that for each construct two indicators must have invariant loadings and intercepts across the groups (Byrne, Shavelson, & Muthén, 1989; Steenkamp & Baumgartner, 1998). It is not necessary that measurement errors can be constrained equal (strict measurement invariance), but if they can be constrained the result is a more parsimonious model. For a thorough treatment of measurement equivalence across groups, we refer to Meredith (1993), Millsap and Meredith (2007), Vandenberg and Lance (2000), and van de Schoot, Lugtig and Hox (2012). For a treatment of longitudinal measurement equivalence, we refer to T. D. Little (2013). In our analysis, we follow the approach outlined in T. D. Little (2013), only omitting the *metric* model because we analyze ordinal categorical items.

In our data, longitudinal measurement equivalence is tested in two steps. The first step is to estimate a model separately for each latent variable in which both time points are included. Note that we tested measurement equivalence separately for each subscale because estimation of the full five factors and two time points model leads to a large and complicated model. The model in the first step is the longitudinal equivalent of the *configural model*, which is the weakest form of factorial invariance. All parameters are freely estimated across waves, what is tested is whether the same factor structure holds for both waves. For the analyses of the model of Step 1 (configural models), the M*plus* default parameterization is used: the first factor loading of a construct will be fixed at 1, and the mean of the latent variables are fixed at 0. The second step is a model where both factor loadings and thresholds are set to be equal across waves. This tests whether the meaning of the construct (factor loadings) and the levels of the underlying items (thresholds) are equal over time, which tests for scalar invariance or strong factorial invariance. For the analyses of the model of Step 2 (constrained model), all factor loadings and thresholds are estimated, and the mean of the latent variable of Wave 1 is fixed at 0, and the variance of the latent variable of Wave 1 is fixed at 1 to identify the model. The mean and variance of the latent variable at Wave 2 are freely estimated.

With this large SDQ data set ($N = 8{,}806$), the chi-square test is expected to indicate lack of measurement equivalence even when there are only very small and negligible differences between the loadings and thresholds across waves. Therefore, the comparative fit index (CFI) difference test is used as an alternative (Cheung & Rensvold, 2002; Meade, Johnson, & Braddy, 2008). Meade et al. (2008) and Cheung and Rensvold (2002) use different cutoff values to decide whether there is invariance. In our analyses, the choice was made to use the cutoff value of .01 from Cheung and Rensvold (2002), based on the observation by Meade et al. (2008) that when there is measurement invariance according to an alternative fit index (e.g., the CFI difference test), and the sample size is larger than 200, it can be assumed there is measurement invariance, even when the chi-square test is significant. With a sample size of more than 8,000 participants, we considered a difference of .01 between the CFI values to be an appropriate cutoff value. For the sake of complete reporting, we present both the significance test and the goodness-of-fit indices. When the two difference tests lead to different conclusions, we base our conclusion on the CFI difference test.

If the model estimated at Step 1 does not fit, there is no Step 2. The following rules of thumb are used to determine the fit of the models. A nonsignificant p value of the chi-square test indicates that the model fits the data. For the root mean square error of approximation (RMSEA) a value $<$ .08 is judged as acceptable, and a value $<$.05 as good. For the CFI and Tucker–Lewis fit index (TLI), a value $>$.90 indicates an acceptable fit, and a value $>$.95 indicates a good

Table 2. Trait and Method Factor Loadings for the MTMM Model.[a]

	Trait factor loadings					Method factor loadings		
	ADHD	COND	EMOT	PEER	PROS	Parent ratings	Teacher ratings	Residual variance
Parent ratings								
ADHD	.753					.254		.368
COND		.686				.361		.398
EMOT			.418			.694		.344
PEER				.613		.333		.513
PROS					−.537	−.135		.693
Teacher ratings								
ADHD	.590						.482	.420
COND		.514					.599	.377
EMOT			.748				.230	.388
PEER				.669			.355	.426
PROS					−.467		−.573	.453

Note. MTMM = multitrait–multimethod approach; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.
[a]Table shows standardized estimates. All estimates are statistically significant ($p < .05$).

fit (Byrne, 2012). The fit of Model 2 (scalar equivalence) is compared with the fit of Model 1 (configural equivalence). A chi-square difference test and CFI difference test will be performed to test the null hypothesis that the fit of both models is equal. When the difference tests are nonsignificant ($\Delta\chi^2$ with $p > .05$, and ΔCFI $\leqslant .01$), it can be assumed that the fit in the constrained model is not significantly worse than the fit of the configural model, and that the factor loadings and thresholds are equal over time.

Results

Convergent and Discriminant Validity

The SDQ instrument has five traits, ADHD, COND, EMOT, PEER, and PROS, which are measured by two methods: parent ratings and teacher ratings. For the MTMM analysis, we use the parent and teacher ratings from the first wave. The measures are constructed as the mean of the item scores for each subscale, where a maximum of one missing value per scale was allowed. The appendix shows the correlations between the 10 measures. Since there was a small amount of missing data, these correlations were estimated using the Expectation–Maximization algorithm in SPSS. The correlations between measures for the same construct based on different methods are given in bold.

The analysis is performed using M*plus* version 7.11 (L. K. Muthén & B. O. Muthén, 1998-2012), using Maximum Likelihood estimation. The MTMM model fits the data well ($\chi^2 = 119.17$, degrees of freedom = 15, TLI = .97, CFI = .99, RMSEA = .04, SRMR = .02). Table 2 presents the estimated factor loadings and residual variances. All factor loadings and variances differ significantly from zero ($p <$ .05). The trait factor loadings are all higher than .4, which puts them in the medium to large range suggested by Cohen (1988) for correlations. For most measures, the trait loadings are larger than the method loadings, but this is not the case for EMOT ratings by parents, and COND and PROS ratings by teachers.

Table 3. Trait Correlations for MTMM Model.[a]

Traits	ADHD	COND	EMOT	PEER	PROS
ADHD	1.000				
COND	.728	1.000			
EMOT	.238	.342	1.000		
PEER	.430	.583	.559	1.000	
PROS	.537	.698	.129	.368	1.000

Note. MTMM = multitrait–multimethod approach; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.
[a]Standardized estimates. All correlations are significant ($p < .05$). Note that the PROS factor loadings in Table 2 are negative, and the latent factor indicates lack of PROS.

Table 3 presents the trait correlations. Most correlations indicate acceptable discriminant validity. The exception is the correlation between ADHD and COND of .75, which is very high, and above the limit of .7 that T. D. Little (2013) prefers.

A more formal way to assess the validity of the measures in the MTMM is to assess the amount of trait, method, and error variance (Alwin, 1974; Scherpenzeel & Saris, 1997; Widaman, 1985). In an MTMM analysis, the square of the factor loadings for the trait factors is interpreted as the validity, and the sum of the squared trait and method loadings indicates the reliability of a measure. Table 4 presents

Table 4. Validity and Reliability of the SDQ Subscales.

	Parent		Teacher	
	Validity	Reliability	Validity	Reliability
ADHD	.567	.632	.348	.580
COND	.471	.601	.264	.623
EMOT	.175	.657	.560	.613
PEER	.376	.487	.448	.574
PROS	.288	.306	.218	.546

Note. SDQ = Strengths and Difficulties Questionnaire; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.

Table 5. Logistic Regression Predicting Attrition at Wave 2.[a]

Predictor	β	*SE*	*OR*
Gender male	.274*	.048	1.315
Home language			
Norwegian	−.299	.233	0.741
Other	.356	.258	1.428
Grade			
5th Grade	−.265*	.059	0.768
6th Grade	−.087	.058	0.917

Note. SE = standard error; *OR* = odds ratio.
[a]Reference categories: Home language: Norwegian and other; Grade: 7th Grade.
*$p < .001$.

the MTMM-based validity and reliability of the SDQ subscales for parents and teachers. It would be nice if we could unequivocally state that either parents or teachers are the best raters, but the results in Table 4 indicate that this is not the case. The reliabilities are below the criterion of .7 suggested by Nunnally (1978). However, the SDQ subscales are short, and they are not self-reports but ratings, so relatively low reliability coefficients may be unavoidable.

Attrition

To analyze the considerable attrition shown in Table 1, we carried out two logistic regressions: one predicting absence on Wave 2 using Wave 1 predictor variables language at home, gender, and grade; and one predicting absence at Wave 1 using wave two predictor variables language at home, gender, and grade at Wave 2.

Table 5 shows the results of the first logistic regression. Gender and grade are significant (based on the Wald criterion, $p < .001$). The different categories of language at home were not significant predictors. Exp(B) values indicate that the odds for attrition for boys are 1.315 times the odds of attrition for girls ($p < .001$), and the odds for attrition for children in the 5th grade are .768 times the odds of attrition for children in the 7th grade ($p < .001$). So, the odds of being absent at Wave 2 are higher for boys compared with girls, and higher for children in the 7th grade compared with children in the 5th grade. However, classification accuracy is not great at 57% classified correctly, and the Nagelkerke pseudo R^2 was .016, which is far below the effect size criteria for a small effect suggested by Zumbo and Thomas (1997) and by Jodoin and Gierl (2001).

The second logistic regression analysis was performed on being absent in Wave 1 as outcome and two predictors: gender and grade. It was not possible to include language at home in this analysis, because for 1,638 participants this information is missing and the participants for whom this information is present, were all present at both waves. A test of the full model with the two predictors against a null-model was statistically nonsignificant, $\chi^2(3, n = 5{,}633) = 2.645$, $p = .450$, and there were also no significant predictors based on the Wald test.

Since attrition was not MCAR, we decided to use weighting on gender and grade to deal with the attrition. The multigroup confirmatory factor analysis models reported in the next section were all estimated with and without weights, and finally using listwise deletion. When the parameter estimates and fit indices of these analyses are compared, we find negligible differences. Therefore, the results reported for the longitudinal measurement equivalence in the next section are unweighted estimates, using Weighted Least Squared estimation, which includes incomplete data in the estimation but still assumes MCAR.

Parent–Teacher and Longitudinal Measurement Equivalence

Parent–Teacher measurement invariance is investigated in Wave 1 and Wave 2 separately. Since our focus is on longitudinal measurement invariance, we summarize the results here. A more detailed report is available online (https://bit.ly/2LhiRCA). In both waves, the configural model fits fine. The full measurement invariance model has a mediocre fit in Wave 1 and an acceptable fit in Wave 2. A good fit is achieved in Wave 1 by freeing one factor loading and one threshold, and in Wave 2 by freeing one factor loading.

Model Tests. The fit of the configural models is acceptable to good. Only for the parent ratings for the construct ADHD, modification indices indicated residual correlations between the Items 2 and 10, and Items 15 and 25. This is acceptable because Items 2 and 10 both refer to hyperactivity, and Items 15 and 25 both to concentration problems. For parsimony, these correlations are not reported here. Table 6 presents the fit information for all models, and the model fit of the full scalar models are in Table 7.

Table 6. Model Fit for Configural Models: All parameters Are Freely Estimated Over Time.[a]

Model	χ^2	*df*	*p*	Scaling factor	CFI	TLI	RMSEA
Parent ratings							
ADHD	1359.688	30	<.001	.5106	.981	.972	.074
COND	563.319	34	<.001	.6564	.954	.939	.044
EMOT	1015.357	34	<.001	.6333	.957	.943	.059
PEER	289.764	34	<.001	.5767	.988	.984	.030
PROS	473.293	34	<.001	.6560	.980	.973	.040
Teacher ratings							
ADHD	1581.470	34	<.001	.5765	.992	.989	.073
COND	213.787	34	<.001	.6386	.990	.987	.025
EMOT	608.701	34	<.001	.6495	.981	.975	.045
PEER	285.089	34	<.001	.5591	.991	.988	.030
PROS	251.806	34	<.001	.5791	.997	.996	.028

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.
[a]In the model for parents for the construct ADHD correlations are included between Items 2 and 10, and between Items 15 and 25.

Table 7. Model Fit for Full Scalar Models: All Factor Loadings and Thresholds Are Constrained to Be Equal Over Time.[a]

Model	χ^2	*df*	*p*	Scaling factor	CFI	TLI	RMSEA
Parent ratings							
ADHD	1625.169	44	<.001	.6545	.978	.977	.066
COND	678.878	48	<.001	.7696	.945	.949	.040
EMOT	1565.947	48	<.001	.7285	.933	.937	.062
PEER	348.022	48	<.001	.7045	.985	.986	.028
PROS	508.138	48	<.001	.7534	.979	.980	.034
Teacher ratings							
ADHD	1634.119	48	<.001	.6923	.992	.992	.063
COND	314.379	48	<.001	.7499	.985	.986	.026
EMOT	677.060	48	<.001	.7449	.980	.981	.039
PEER	311.397	48	<.001	.6877	.990	.991	.025
PROS	403.149	48	<.001	.6889	.995	.995	.030

Note. df = degrees of freedom; CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.
[a]In the model for parents for the construct ADHD correlations are included between Items 2 and 10, and between Items 15 and 25.

Difference Tests. The next step is to investigate whether there is a significant difference between the full scalar models and the configural models. For this test, the Satorra–Bentler scaled chi-square difference test and CFI difference test are used. A nonsignificant difference ($\Delta\chi^2 p > .05$, $\Delta\text{CFI} \leqslant .01$) indicates that the fit of the model with constrained factor loadings and thresholds is not significantly worse than the configural model. When the chi-square difference test was significant, indicating there is no full scalar invariance, we continue investigating partial scalar invariance. Based on the modification indices, we decided which thresholds and factor loadings needed to be freed over time. This step was repeated until (a) a nonsignificant chi-square different test statistic was found, (b) when the thresholds and factor loadings of no more than two items were constrained, or (c) when there were no alternatives indicated by the modification indices.

Tables 8 and 9 present the results for parent ratings and teacher ratings, respectively. Model 1 represents for all constructs the full scalar model: The thresholds and factor loadings of all five items are constrained to be equal over time. In the subsequent models, the thresholds and factor loadings of one item at a time were unconstrained which leads to partial scalar equivalence.

In Tables 8 and 9, it can be seen that based on the chi-square test, none of the constructs have full scalar invariance and only for a few of the models there is partial scalar invariance (parent ratings: PEER and PROS; teacher ratings: ADHD and COND). Based on the CFI difference test, there is full scalar invariance for all constructs (for parents and teacher ratings). As aforementioned, because of the sample size dependency of the chi-square test, the fit of the constrained models and the CFI difference test is leading in the decision about measurement invariance.

Discussion

The main aims of the present study were to assess if parents and teachers measured the same construct for youth mental health when using the SDQ, and whether there was measurement equivalence over time. The results of the MTMM analyses are encouraging: the MTMM model fits well, and there is evidence for both convergent and discriminant validity. The validity and reliability of the SDQ scales are modest, in line with previous investigations (Hill & Hughes, 2007). Given that these are short scales, used in a survey type of data collection, low reliabilities are expected (Heath & Martin, 1997). Increasing the reliability by including more items seems desirable, but it could lead to a scale that is too long for use in a survey context. The MTMM analysis also indicates clear method effects, also found in previous investigations (Hill & Hughes, 2007). All of this points to the following two recommendations. First, it is good practice to collect SDQ observations from both parents and teachers and to combine them. This is in line with clinical practice, as for some mental health problems, such as symptoms of ADHD, it is an additional requirement to collect ratings across multiple contexts, to cover the information required by diagnostic criteria. Second, given the modest reliability and validity, it is good practice to use latent variable modeling and not observed (sum) scores when carrying out substantive research with the SDQ.

Table 8. $\Delta\chi^2$ and ΔCFI for Parent Ratings: Comparison of the (Partial) Scalar and Configural Models.[a]

Model	Constraints on thresholds and factor loadings of items	$\Delta\chi 2$ (*df*), *p*	ΔCFI
ADHD			
1	2, 10, 15, 21, 25	383.67 (14), <.001	−.003
2	2, 10, 15, 25	244.44 (11), <.001	−.001
3	2, 15, 25	143.91 (8), <.001	.000
4	2, 25	85.74 (5), <.001	.001
COND			
1	5, 7, 12, 18, 22	146.19 (14), <.001	−.009
2	5, 7, 12, 22	81.47 (11), <.001	−.003
3	5, 7, 22	44.19 (8), <.001	.000
4	5, 7	25.57 (5), <.001	.000
EMOT			
1	3, 8, 13, 16, 24	518.67 (14), < .001	−.006
2	3, 8, 13, 24	296.81 (11), <.001	.001
3	3, 8, 13	158.7 (8), <.001	.008
4	8, 13	64.81 (5), <.001	.012[b]
PEER			
1	6, 11, 14, 19, 23	76.93 (14), <.001	−.003
2	11, 14, 19, 23	26.58 (11), .019	.000
3	11, 19, 23	12.99 (8), .112	.001
PROS			
1	1, 4, 9, 17, 20	73.09 (14), <.001	−.001
2	4, 9, 17, 20	46.95 (11), <.001	.000
3	9, 17, 20	20.15 (8), .103	.001

Note. CFI = comparative fit index; *df* = degrees of freedom; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.
[a]$\Delta\chi^2$ with $p > .05$ and $\Delta CFI \leq .01$ indicate that the constrained model does not fit significantly worse than the configural model. [b]$\Delta CFI > .01$ indicates that the CFI became significantly better.

The high correlation between ADHD and conduct problems could be problematic in relation to discriminant validity between the two subscales. A strong correlation is expected given the high rate of co-occurrence between ADHD and conduct problems in children (Heiervang et al., 2007, Kessler et al., 2005) which may suggest genetic covariance in the two disorders (Dick, Viken, Kaprio, Pulkkinen, & Rose, 2005). The strong association between the subscales has in fact previously motivated researchers to combine the two subscales and use them as a broader externalizing problems scale (Goodman, Lamping, & Ploubidis, 2010). There is also a negative association between prosocial and conduct problems. Conduct problems are characterized by negative behavior which also is related to social relationships and is known to be inversely related to social competence and prosocial behavior (Edwards & Bromfield, 2009). These overlapping subscales also reflect the general high co-occurrence of mental health problems in childhood (Heiervang et al., 2007).

The recommendation based on the MTMM results about using latent variable modelling for SDQ, to deal with the relatively low reliabilities, is also supported by the longitudinal measurement equivalence analysis. In longitudinal data, measurement equivalence is often expected, because we have data on the same subjects using the same instrument. In our case, when the fit indices are used to select a model, full scalar equivalence can be assumed for all constructs. Moreover, releasing constraints to allow partial equivalence results in minimal changes in overall fit. So, the use of observed instead of latent scores appears justifiable (Steinmetz, 2013). However, if we use the chi-square difference test as a criterion, the improvement in going from full scalar equivalence to partial scalar equivalence is significant. This indicates that latent scores are preferred, at least when combining parent and teacher observations.

The attrition analysis shows that demographic background variables have only a small effect on attrition. However, the attrition rate itself is high, and we, therefore, recommend that substantive analyses on such data use modern missing data methods (cf. Enders, 2010) to move the relevant assumption from MCAR toward MAR.

Among the strengths of the current study are the large population-based sample, the relatively large cohort available for longitudinal analyses, the use of multiple raters, and the use of state-of-the-art methods for investigating interrater convergence and longitudinal measurement

Table 9. $\Delta\chi^2$ and ΔCFI for Teacher Ratings: Comparison of the (Partial) Scalar and Configural Models.[a]

Model	Constraints on thresholds and factor loadings of items	$\Delta\chi^2$ (*df*), *p*	ΔCFI
ADHD			
1	2, 10, 15, 21, 25	225.55 (14), <.001	.000
2	2, 15, 21, 25	128.72 (11), <.001	.000
3	2, 15, 25	64.94 (8), <.001	.000
4	15, 25	10.63 (5), .060	.000
COND			
1	5, 7, 12, 18, 22	97.26 (14), <.001	−.005
2	5, 7, 18, 22	30.04 (11), .002	.000
3	5, 18, 22	13.33 (8), .112	.001
EMOT			
1	3, 8, 13, 16, 24	111.60 (14), <.001	−.001
2	8, 13, 16, 24	49.67 (11), <.001	.000
3	13, 16, 24	15.87 (8), .044	.001
PEER			
1	6, 11, 14, 19,23	69.85 (14), <.001	−.001
2	6, 11, 19,23	47.04 (11), <.001	.000
3	6, 19,23	35.23 (8), <.001	.000
4	19,23	26.37 (5), <.001	.000
PROS			
1	1, 4, 9, 17, 20	138.04 (14), <.001	−.002
2	1, 4, 9, 20	81.99 (11), <.001	−.001
3	1, 9, 20	41.34 (8), <.001	.000
4	9, 20	21.21 (5), <.001	.000

Note. CFI = comparative fit index; *df* = degrees of freedom; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior.
[a]$\Delta\chi^2$ with $p > .05$ and ΔCFI ⩽ .01 indicate that the constrained model does not fit significantly worse than the configural model.

equivalence. However, the results must also be interpreted in light of some limitations. One issue regards the relatively high attrition from Wave 1 to Wave 2, although this was partly addressed in the analyses. Another issue regards the relatively narrow age-span included in the current study, which limits our ability to generalize the findings from the current study to broader age-groups. Further research should include the third and fourth waves of the Bergen Child Study, to investigate whether the results of longitudinal measurement equivalence for parents and teachers also hold in these waves. Additionally, it would be of interest to see if there is measurement equivalence over time in children's self-reports, which were introduced in the second wave of the study. Furthermore, additional studies could investigate how children's self-reports relate to the parent and teacher reports within the same wave. On a more general note, there are also relevant questions related to how strong the interrater convergence can be expected to be, given that children operate in different surroundings influencing the behavior they express in these distinct contexts.

In conclusion, the results of the present study indicate good convergent validity for parent and teacher ratings of the SDQ—a frequently used screening measure for youth mental health. However, we also find considerable method variance across raters, and relatively low reliability and validity for some of the SDQ subscales. The first recommendation from the current study is to use latent variable modeling instead of sum scores, in substantive assessments of interrater convergence and for longitudinal modeling with the SDQ. Second, the results support the clinical practice and recommendations of collecting and combining parent and teacher ratings for the assessment of youth mental health. The results from the current study suggest this would enhance both reliability and validity.

Appendix

Correlations Between the 10 Means of the Subscales: Standardized Estimates Using ML Estimation in M*plus*.

	Parents					Teachers				
r	ADHD1	COND1	EMOT1	PEER1	PROS1	ADHD1	COND1	EMOT1	PEER1	PROS1
Parents										
ADHD1	1									
COND1	.478	1								
EMOT1	.268	.351	1							
PEER1	.297	.365	.367	1						
PROS1	−.240	−.335	−.120	−.170	1					
Teachers										
ADHD1	**.454**	.298	.104	.230	−.141	1				
COND1	.282	**.352**	.118	.270	−.158	.514	1			
EMOT1	.132	.156	**.307**	.231	−.038	.211	.285	1		
PEER1	.205	.231	.156	**.425**	−.097	.325	.426	.390	1	
PROS1	−.210	−.220	−.070	−.160	**.233**	−.430	−.498	−.187	−.334	1

Note. ML = Maximum Likelihood; ADHD = hyperactivity–inattention problems; COND = conduct problems; EMOT = emotional symptoms; PEER = peer relationship problems; PROS = prosocial behavior. For all correlations $p < .05$. The correlations between measures for the same construct based on different methods are given in bold.

Declaration of Conflicting Interests

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

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Sanne C. Smid was supported by a grant from the Netherlands Organization for Scientific Research, NWO-VIDI-452-14-006.

Supplemental Material

Supplemental material for this article is available online.

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.

Alwin, D. (1974). An analytic comparison of four approaches to the interpretation of relationships in the multitrait–multimethod matrix. In H. L. Costner (Ed.), *Sociological methodology, 1973-1974* (pp. 79-105). San Francisco, CA: Jossey-Bass.

Becker, A., Rothenberger, A., Sohn, A., & BELLA Study Group. (2015). Six years ahead: A longitudinal analysis regarding course and predictive value of the Strengths and Difficulties Questionnaire (SDQ) in children and adolescents. *European Child & Adolescent Psychiatry*, *24*, 715-725.

Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): Factor structure and gender equivalence in Norwegian adolescents. *PLoS ONE*, *11*, e0152202. doi:10.1371/journal.pone.0152202

Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. New York, NY: Routledge.

Byrne, B. M., Shavelson, R. J., & Muthén, B. O. (1989). Testing for the equivalence of factor and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105*, 456- 466.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105.

Cheng, S., Keyes, K. M., Bitfoi, A., Carta, M. G., Koç, C., Goelitz, D., . . . Kovess-Masfety, V. (2018). Understanding parent–teacher agreement of the Strengths and Difficulties Questionnaire (SDQ): Comparison across seven European countries. *International Journal of Methods in Psychiatric Research*, *27*, e1589.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling*, *9*, 233-255.

Chiorri, C., Hall, J., Casely-Hayford, J., & Malmberg, L.-E. (2016). Evaluating measurement invariance between parents using the Strengths and Difficulties Questionnaire (SDQ). *Assessment*, *23*, 63-74. doi:10.1177/1073191114568301

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Academic Press.

Costello, E. J., Foley, D. L., & Angold, A. (2006). 10-Year research update review: The epidemiology of child and adolescent psychiatric disorders: II. Developmental epidemiology. *Journal of the American Academy of Child & Adolescent Psychiatry*, *45*, 8-25. doi:10.1097/01.chi.0000184929.41423.c0

DeVries, J. M., Gebhardt, M., & Voß, S. (2017). An assessment of measurement invariance in the 3-and 5-factor models of the Strengths and Difficulties Questionnaire: New insights from a longitudinal study. *Personality and Individual Differences*, *119*, 1-6.

Dick, D. M., Viken, R. J., Kaprio, J., Pulkkinen, L., & Rose, R. J. (2005). Understanding the covariation among childhood externalizing symptoms: Genetic and environmental influences on conduct disorder, attention deficit hyperactivity dis-

order, and oppositional defiant disorder symptoms. *Journal of Abnormal Child Psychology*, *33*, 219-229.

Edwards, B., & Bromfield, L. M. (2009). Neighborhood influences on young children's conduct problems and pro-social behavior: Evidence from an Australian national sample. *Children and Youth Services Review*, *31*, 317-324.

Enders, C. K. (2010). *Applied missing data analysis*. New York, NY: Guilford Press.

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*, 581-586.

Goodman, R. (1999). The extended version of the Strengths and Difficulties Questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of Child Psychology and Psychiatry*, *40*, 791-799.

Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2000). Using the Strengths and Difficulties Questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *British Journal of Psychiatry*, *177*, 534-539.

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

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.

Heath, A., & Martin, J. (1997). Why are there so few formal measuring instruments in social and political research? In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz & D. Trewin (Eds.), *Survey measurement and process quality* (pp. 71-86). New York, NY: Wiley.

Heiervang, E., Stormark, K. M., Lundervold, A. J., Goodman, R., Posserud, M.-B., & Ullebø, A. K. (2007). Psychiatric disorders in Norwegian 8- to 10-year-olds. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*, 438-447. doi:10.1097/chi.0b013e31803062bf

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

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating Type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, *14*, 329-349.

Jöreskog, K. G. (1971). Statistical analysis of congeneric tests. *Psychometrika*, *36*, 109-133.

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Merikangas, K. R., & Walters, E. E. (2005). Lifetime prevalence and age-of-onset distributions of *DSM-IV* disorders in the National Comorbidity Survey replication. *Archives of General Psychiatry*, *62*, 593-602. doi:10.1001/archpsyc.62.6.593

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York, NY: Wiley.

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

Marsh, H., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. Hoyle (Ed.), *Structural equation modeling* (pp. 177-198). Thousand Oaks, CA: Sage.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology*, *93*, 568-592.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127-143.

Meredith, W. (1993). Measurement invariance, factor-analysis and factorial invariance. *Psychometrika 58*, 525-543.

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

Millsap, R. E., & Meredith, W. (2007). Factorial invariance: Historical perspectives and new problems. In R. Cudeck & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 131-152). Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479-515.

Munkvold, L., Lundervold, A., Lie, S. A., & Manger, T. (2009). Should there be separate parent and teacher-based categories of ODD? Evidence from a general population. *Journal of Child Psychology and Psychiatry*, *50*, 1264-1272. doi:10.1111/j.1469-7610.2009.02091.x

Muthén, B. O. (2013). *Measurement invariance with multigroups*. Retrieved from http://www.statmodel.com/discussion/messages/9/11980.html?1456792993

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

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

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

Rønning, J. A., Handegaard, B. H., Sourander, A., & Mørch, W. T. (2004). The Strengths and Difficulties Self-Report Questionnaire as a screening instrument in Norwegian community samples. *European Child & Adolescent Psychiatry*, *13*, 73-82.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*, 581-592.

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

Sayal, K., Heron, J., Golding, J., & Emond, A. (2007). Prenatal alcohol exposure and gender differences in childhood mental health problems: A longitudinal population-based study. *Pediatrics*, *119*, e426-434. doi:10.1542/peds.2006-1840

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*, 147-177.

Scherpenzeel, A. C., & Saris, W. E. (1997). The validity and reliability of survey questions. A meta-analysis of MTMM studies. *Sociological Methods & Research*, *25*, 341-383.

Sosu, E. M., & Schmidt, P. (2017). Tracking emotional and behavioral changes in childhood: Does the Strength and

Difficulties Questionnaire measure the same constructs across time? *Journal of Psychoeducational Assessment*, *35*, 643-656.

Steenkamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 1, 78-90.

Steinmetz, H. (2013). Analyzing observed composite differences across groups. Is partial measurement invariance enough? *Methodology*, *9*, 1-12.

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

Stormark, K. M., Heiervang, E., Heimann, M., Lundervold, A., & Gillberg, C. (2008). Predicting nonresponse bias from teacher ratings of mental health problems in primary school children. *Journal of Abnormal Child Psychology*, *36*, 411-419. doi:10.1007/s10802-007-9187-3

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman and Hall/CRC.

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 de Schoot, R., Lugtig, P., & Hox, J. J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-492.

Vandenberg, R., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4-70.

Widaman, K. F. (1985). Hierarchically nested covariance structure models for multitrait-multimethod data. *Applied Psychological Measurement*, *9*, 1-26.

Zumbo, B. D., & Thomas, D. R. (1997). *A measure of effect size for a model-based approach for studying DIF* (Working Paper). Prince George, British Columbia, Canada: University of Northern British Columbia.