Article

A Longitudinal and Gender Invariance Analysis of the Strengths and Difficulties Questionnaire Across Ages 3, 5, 7, 11, 14, and 17 in a Large U.K.-Representative Sample

Assessment
2022, Vol. 29(6) 1248–1261

DOI: 10.1177/10731911211009312
journals.sagepub.com/home/asm

Aja Louise Murray[1], Lydia Gabriela Speyer[1], Hildigunnur Anna Hall[1], Sara Valdebenito[2], and Claire Hughes[2]

Abstract

Developmental invariance is important for making valid inferences about child development from longitudinal data; however, it is rarely tested. We evaluated developmental and gender invariance for one of the most widely used measures of child mental health: the parent-reported Strengths and Difficulties Questionnaire (SDQ). Using data from the large U.K. population-representative Millennium Cohort Study ($N = 10{,}207$; with data at ages 3, 5, 7, 11, 14, and 17 years), we tested configural, metric, scalar, and residual invariance in emotional problems, conduct problems, hyperactivity/inattention, prosociality, and peer problems. We found that the SDQ showed poor fit at age 3 in both males and females and at age 17 in males; however, it fit reasonably well and its scores were measurement invariant up to the residual level across gender at ages 5, 7, 11, and 14 years. Scores were also longitudinally measurement invariant across this age range up to the partial residual level. Results suggest that the parent-reported SDQ can be used to estimate developmental trajectories of emotional problems, conduct problems, hyperactivity/inattention, prosociality, and peer problems and their gender differences across the age range 5 to 14 years using a latent model. Developmental differences outside of this range may; however, partly reflect measurement differences.

Keywords

longitudinal measurement invariance, gender measurement invariance, strengths and difficulties questionnaire

The majority of mental health problems and related psychosocial impairments begin before adulthood (Davies, 2013). From a prevention perspective there is, therefore, considerable interest in tracking developmental trajectories of mental health from the earliest stages of life in order to characterize normative versus nonnormative trajectories and to establish the risk and protective factors that distinguish them (Parkes et al., 2016; Patalay et al., 2017). A core challenge in these efforts relates to measuring symptoms and impairments in a comparable manner across different developmental stages (Knight & Zerr, 2010; Murray, Obsuth, et al., 2017). Given that gender differences are evident in most common mental health issues both in terms of their levels and developmental trajectories, there is also a need to ensure comparability of scores across gender (Booth & Murray, 2018). The issue of longitudinal measurement invariance has received relatively little attention, especially in combination with gender differences; however, valid inferences regarding development from longitudinal studies critically depend on it. In this study, we, therefore, examine the longitudinal and gender measurement invariance of one of the most widely used measure of child psychopathology: the Strengths and Difficulties Questionnaire (SDQ; Goodman, 1997) in a large nationally representative sample of children in the U.K. measured from age 3 to age 17.

Longitudinal studies of child and adolescent development represent an invaluable resource for illuminating developmental trajectories of mental health issues from the very beginning of life (e.g., Niarchou et al., 2015). When analyzed with methods such as growth curve or growth

[1]University of Edinburgh, Edinburgh, UK
[2]University of Cambridge, Cambridge, UK

Corresponding Author:
Aja Louise Murray, Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK.
Email: aja.murray@ed.ac.uk

mixture models, longitudinal data provide critical information about normative symptom trajectories, variation around normative trajectories, and predictors and outcomes of those trajectories (Murray, Eisner, et al., 2017; Parkes et al., 2016; Patalay et al., 2017; Speyer et al., 2020). When analyzed using methods such as (random intercept) cross-lagged panel models or autoregressive latent trajectory models with structured residuals, they provide information about the developmental relations between symptoms and factors such as peer and academic problems, parent conflict, and symptoms of other disorders (Besemer et al., 2016; Murray, Eisner, et al., 2019; van Lier et al., 2012; Wertz et al., 2015).

However, valid inferences from these methods depend on measuring symptoms in a comparable manner over development (Edwards & Wirth, 2012). In fact, given the widespread biological and social–environmental changes that occur over and between childhood and adolescence, it is quite plausible that the meaning and manifestation of symptoms could differ across developmental stages (e.g., Rapee et al., 2019). For example, it has been suggested that hyperactivity symptoms may have a different manifestation over development, with overt behaviors (e.g., difficulty sitting down for a prolonged period of time) giving way to internal feelings of restlessness at later stages of development (Weyandt et al., 2003). Similarly, whereas physical aggression is relatively common and quite normal in the preschool years, it becomes a marker of much more severe problems by the time of late adolescence (Nærde et al., 2014; Tremblay, 2000).

Similar considerations apply to gender differences. Males and females have previously shown differences in levels and developmental trajectories of many mental health problems, with later onset and emotional problems tending to be more common in females and earlier onset and neurodevelopmental and behavioral issues tending to be more common in males (e.g., Martel, 2013; Rutter et al., 2003). Whenever there are differences in the prevalence of mental health issues across males and females there are concerns that symptoms may be more difficult to detect in the gender with the lower prevalence because informants are less attuned to symptoms in these cases. More than this, however, there is also evidence that the manifestation of mental health issues can differ across males and females. In attention deficit hyperactivity disorder, for example, it has been proposed that females may show profiles that are more characterized by inattention (as opposed to hyperactivity) and emotional symptoms as compared to males (Gershon & Gershon, 2002; Williamson & Johnston, 2015). Similarly, in relation to conduct problems, physical forms of aggression tend to be relatively more common in males than females, compared with social aggression, which shows little to no gender difference (Archer, 2004).

These developmental and gender differences are likely to be manifested as violations of measurement invariance over time, or "longitudinal measurement invariance' and across gender or 'gender measurement invariance' (Edwards & Wirth, 2012). Measurement invariance refers to the distribution of item scores given the level of an underlying latent trait being independent of the measurement occasion (for longitudinal invariance) or group (e.g., gender invariance; Millsap, 2012). That is, after taking into account changes in latent trait levels across time/group, a person's scores should not depend on whether they came from Wave 1, 2, 3, and so on, or were male or female. Full measurement invariance is difficult to test because of the need to demonstrate that the entire distribution of scores is the same across time/group conditional on latent trait level (see Molenaar & Borsboom, 2013, for a discussion). However, a weaker version: factorial invariance is readily testable within a confirmatory factor analysis (CFA) framework and is often used to test whether a measure shows any evidence of differential functioning across time or group (Liu et al., 2017; Van de Schoot et al., 2012). In a CFA framework, observed variables (e.g., questionnaire items) are defined as indicators of underlying latent factors, such as "internalising problems" or "externalizing problems" (Millsap & Yun-Tein, 2004). Observed indicators are linked to the latent factors by factor loadings that represent the strength of association between indicators and factors. In addition, for the ordinal items which are commonly used in questionnaires (e.g., Likert-type items), thresholds are modelled and are the point on the latent factor scale that individuals transition from scoring in one category (e.g., "strongly disagree") to the next category (e.g., "disagree"; Millsap & Yun-Tein, 2004).

Different levels of invariance are required for different kinds of inferences from longitudinal data using ordinal indicators (Liu et al., 2017). To compare variances and regression paths involving latent factors over time/group (e.g., to test whether a predictive path is the same over different stages of development) requires "metric invariance," which is when the factor loadings are equal across time/group. To compare factor means over time/group (e.g., in a growth curve model), "scalar invariance" is required, which is when both loadings and thresholds are equal across time/group (Liu et al., 2017; Liu & West, 2018). If this does not hold then true change over time may be masked or overstated. To compare means over time based on observed scores (as opposed to based on latent factor means) requires "residual invariance," where residual variances are also equal over time/group (Liu et al., 2017).

Where measurement invariance does not hold at a given level, it is often possible to obtain a partially invariant model which is sufficient to make variance/covariance/mean comparisons over time based on the latent factors, provided that the noninvariance is appropriately modelled (Edwards & Wirth, 2012; Pokropek et al., 2019). The more items that show invariance the better, however, simulation studies suggest that up to 80% of the items can be noninvariant without

necessarily invalidating inferences made regarding structural parameters (Pokropek et al., 2019).

Despite its importance, relatively limited attention has been paid to the issue of longitudinal invariance in child and adolescent mental health. Only a relatively small number of studies have tested the longitudinal invariance of mental health instruments used in child and adolescent studies (Croft et al., 2015; Leopold et al., 2016; Mathyssek et al., 2013; Motl et al., 2005; Murray, Obsuth, et al., 2017; Sosu & Schmidt, 2017; Sterba et al., 2010; Verhoeven et al., 2013). One of the most commonly used omnibus measures of child psychopathology is the SDQ (Goodman, 1997). It is widely used both in longitudinal studies and in clinical and educational settings, where it informs the provision of services and is used to evaluate the effects of interventions. In some countries, such as Scotland and the Netherlands, it is employed to track the impact of child-related policies (Sosu & Schmidt, 2017). It measures child psychopathology in five dimensions: hyperactivity/inattention, conduct problems, emotional problems, peer problems, and prosociality. It is available in self-, teacher-, and parent-report versions and is recommended for use across an age range of 3 to 16 years.

The SDQ has undergone extensive psychometric evaluation with numerous studies reporting on its factor structure, reliability, convergent validity, content validity, discriminative validity, and interrater agreement (see review by Kersten et al., 2016). Some studies have also examined its measurement invariance across factors such as gender, informant, and country (Bøe et al., 2016; Ortuno-Sierra et al., 2015; Rogge et al., 2018). However, despite its popularity in longitudinal studies, we could identify only two studies that evaluated the longitudinal measurement invariance of the English language SDQ (Croft et al., 2015; Sosu & Schmidt, 2017). Croft et al. (2015) examined the longitudinal variance of the parent-reported version of the SDQ in the Millennium Cohort Study (MCS; Connelly & Platt, 2014) across ages 3, 5, and 7 years. They found that both metric and scalar measurement invariance held for the conduct problems, hyperactivity/inattention and prosocial behavior subscales; however, only metric measurement invariance held for the emotional and peer problems scales. More recently, Sosu and Schmidt (2017) examined the longitudinal measurement invariance of the parent-reported SDQ in the *Growing Up in Scotland* study (https://growingupinscotland.org.uk/) across ages 4, 5, and 6 years. They found that metric and scalar measurement invariance was supported for all five subscales over time. No study has yet examined the measurement invariance of the SDQ from childhood into and across adolescence. This is an important gap because substantial physical, cognitive, social, and biological changes occur between childhood and adolescence and within adolescence itself that could undermine the developmental comparability of measures of psychopathology (Rapee et al., 2019). Indeed, in the transition to adolescence the risk for the onset of several disorders increases, especially emotional disorders (Copeland et al., 2014; Roza et al., 2003). Furthermore, despite the fact that gender differences in developmental trajectories of mental health problems are also commonly of interest (Cleverley et al., 2012; Dekker et al., 2007; Murray, Booth, et al., 2019; Salk et al., 2016), no study has yet examined developmental measurement invariance alongside gender measurement invariance for the SDQ scores. Among the few studies that have examined gender measurement invariance in versions of the SDQ, some have identified violations of measurement invariance (Bøe et al., 2016; van de Looij-Jansen et al., 2011); however, it is not clear how these violations interact with developmental stage as longitudinal and gender measurement invariance are yet to be tested together.

It is not, therefore, clear whether gender differences in developmental trajectories and their relations across dimensions and to risk factors and outcomes can be validly compared and/or whether males and females can be validly combined into a single sample (see, e.g., Meredith & Teresi, 2006). Here, we address these gaps by evaluating the gender and longitudinal invariance of the parent-reported SDQ across ages 3, 5, 7, 11, 14, and 17 years using the large, population-representative MCS.

Method

Participants

Participants ($N = 10{,}207$) were children/adolescents from U.K.-based the MCS (Connelly & Platt, 2014) who participated up to age 17 (the initial sample was 18,818). The MCS is a longitudinal study that has tracked the family and broader social lives of children born at the beginning of the 21st century (2000-2002). Families were sampled using a stratified sampling procedure and included participants from all four nations of the United Kingdom. The sampling frame was families with a 9-month-old child and eligible for the (universal) child benefit at the time of the first sweep (measurement wave). They were identified based on *Department for Social Security* (now *Department for Work and Pensions*) Child Benefit register records. Sensitive cases (including for example where children had died or been taken into local care or where a family were being investigated for benefit fraud) were excluded. A further exclusion criterion was that if a family had already participated in the *Department for Work and Pension's Family and Children Survey.* A small number of families (56) were added who were not initially identified via the Child Benefit register but who could be identified via health visitors. Oversampling of ethnic minority groups and disadvantaged families was built into the survey design with design weights used to correct for this oversampling.

Stratification variables and attrition weights that include an adjustment for baseline study design are included as part

Table 1. Sample Demographic Information.

Variable	Sweep	*N* with SDQ data at each wave	*M*	*SD*
Age	Sweep 2	8,955	3.13	0.19
	Sweep 3	9,371	5.21	0.24
	Sweep 4	9,171	7.23	0.24
	Sweep 5	9,349	10.66	0.48
	Sweep 6	9,076	13.76	0.45
	Sweep 7	8,933	16.68	0.48
		Category	%	*N*
Sex	Sweep 2	Female	50.25	4,500
	Sweep 2	Male	49.75	4,455
	Sweep 3	Female	49.89	4,675
	Sweep 3	Male	50.11	4,696
	Sweep 4	Female	50.22	4,606
	Sweep 4	Male	49.88	4,565
	Sweep 5	Female	50.41	4,713
	Sweep 5	Male	49.59	4,636
	Sweep 6	Female	50.07	4,553
	Sweep 6	Male	49.93	4,523
	Sweep 7	Female	50.07	4,473
	Sweep 7	Male	49.93	4,460
Child ethnicity		White	82.94	7,500
		Other Ethnicity	17.06	1,543
Maternal academic qualification		Higher degree	4.34	392
		First degree	16.67	1,506
		Diplomas in higher education	9.74	880
		A/AS/S Levels	10.43	942
		O level/GCSE Grades A-C	32.44	2,931
		GCSE Grades A-C	9.05	818
		Other academic qualification	2.78	251
		None of these qualifications	14.56	1,316
Deprivation		Most deprived decile	12.98	1128
		10% to <20%	11.78	1,023
		20% to <30%	11.18	971
		30% to <40%	9.67	840
		40% to <50%	9.39	816
		50% to <60%	9.01	783
		60% to <70%	8.28	719
		80% to <90%	9.34	811
		Least deprived decile	9.96	865

Note. These are unweighted and based on the sample of participants with Strengths and Difficulties Questionnaire data up to age 17.

of the data release to account for the complex sampling design. This allows parameter estimates to be corrected for nonrandom sampling and attrition and for standard errors to be corrected for clustering. We used a pseudo-maximum likelihood estimation technique to make these adjustments. Participants were sampled from a population that was born over a 16-month period, which allowed season effects to be taken into account whilst also making the fieldwork more feasible. The England and Wales samples were born between 1st September 2000 and 31st August 2001; the Scotland and Northern Ireland samples were born between 24th November, 2000, and 11th January, 2002. MCS is fully documented at: https://ukdataservice.ac.uk where data can also be downloaded. Seven sweeps of data are currently available, at ages 9 months, 3, 5, 7, 11, 14, and 17 years with parent-reported SDQ data available at Sweeps 2 to 6 (ages 3-17 years).

Sample demographic information is provided in Table 1. These are based on unweighted data from children with SDQ scores at age 17.

Measures

Parent-Reported Strengths and Difficulties Questionnaire. The parent-reported SDQ measures five dimensions of child psychosocial functioning in five subscales: conduct problems, hyperactivity/inattention, emotional problems, peer problems, and prosocial behavior. Each subscale has five items. The conduct problem items refer to: *often having temper tantrums; generally being obedient; often fighting with or bullying other children; often lying or cheating; and stealing from home, school, or elsewhere*. The hyperactivity/inattention items refer to: *being restless, overactive, being unable to stay still for long; constantly fidgeting or squirming; being easily distracted; thinking before acting; and seeing tasks through to their end*. The emotional problems items refer to: *often complaining of headaches, stomach-aches, or sickness; having many worries; being often unhappy, downhearted, or tearful; being nervous or clingy in new situations; and having many fears, being easily scared*. The prosociality items refer to: *being considerate of others' feelings; sharing readily with other children; being helpful if someone is hurt, upset, or feeling ill; being kind to younger children; and often volunteering to help others*. The peer problems items refer to: *being rather solitary and tending to play alone; having at least one good friend; being picked on or bullied by other children; and getting on better with adults than other children*. The SDQ version administered at age 3 was adapted slightly to improve its age-appropriateness. Specifically, in the conduct problems subscale: *argumentative with adults* and *can be spiteful* were used instead of *often lies or cheats*, and *steals from home, school, or elsewhere*; and in the hyperactivity/inattention subscale: *can stop and think before acting* was used instead of *thinks things out before acting*. Responses are recorded on a 3-point scale from *not true* to *certainly true*. Respondents could also select a "can't say" or "not applicable" option.

As noted earlier, there is an extensive literature on the psychometric properties of the SDQ. Most studies support the structural and convergent validity of the five-dimensional model implied by the design of the scale. Internal consistency values of the five subscales based on Cronbach's α have sometimes been noted to be weak; however, in the current study all internal consistency values based on McDonald's (1999) omega calculated using polychoric item correlations to account for their ordered-categorical response format were good, with only one subscale (peer problems at Sweep 2) showing values $<.70$. For emotional problems at Sweeps 2 to 7 (age 3 to 17) omega values were as follows: .76, .78, .80, .83, .83, and .83. For conduct problems they were as follows: .79, .76, .81, .84, .85, and .80. For hyperactivity/inattention problems they were as follows: .80, .84, .86, .86, .85, and .81. For prosociality they were as follows: .77, .80, .83, .82, .86, and .82. For peer problems they were as follows: .65, .70, .75, .80, .78, and .72.

Longitudinal and Gender Measurement Invariance. Prior to conducting measurement invariance analyses, CFAs were fit for each gender individually at each time point. For these and all subsequent analyses, we used a five-dimensional model of the SDQ for our analyses, where each subscale (hyperactivity/inattention, conduct problems, emotional problems, peer problems, and prosociality) mapped to a latent factor. This structure was preferred from among at least 12 different factor structures that have been proposed for the SDQ (Gomez & Stavropoulos, 2019). One of the main alternatives to the five-dimensional model that has been proposed is a three-dimensional model with latent factors for externalizing problems, internalizing problems, and prosociality; however, this has received less empirical support (e.g., Croft et al., 2015). Bifactor and second-order models of the SDQ have also been tested and have often shown good fit; however, de la Cruz et al. (2018) directly compared these models to a five-dimensional model and found the latter to be better fitting. Finally, a model similar to the five-dimensional model but with an additional methods factor ("positive construal") has been proposed (Palmieri & Smith, 2007). Gomez and Stavropoulos (2019) found that this model fit better than a five-dimensional model in their Malaysian sample; however, other studies have failed to find support for this model over the original five-dimensional model (Sanne et al., 2009). On balance, therefore, the five-dimensional model was preferred to other structures that have been proposed because it reflects the theoretical structure for the SDQ and has generally been supported by factor analytic work, including in direct comparisons with alternative structures (e.g., Bøe et al., 2016; de la Cruz et al., 2018; Gomez & Stavropoulos, 2019).

Cross-group (gender) and longitudinal measurement invariance were tested in sequence. We began by testing invariance across gender in each time point. A "forwards" method beginning with the least constrained model and successively adding more constraints was used (see, e.g., Kim & Willson, 2014). Within each time point we began by fitting a configural model (based on the five-dimensional model discussed above) in which the pattern of item loadings was the same across all time but the magnitude of loadings (and thresholds) were allowed to vary across group.

For scaling and identification purposes, the means and variances of all latent factors in one gender (females served as the reference group) were fixed to 0 and 1, respectively, and the loading of one item per latent factor was fixed equal across group, one threshold per item was fixed equal across group, and the second threshold for the marker variable ("reference indicator") was fixed equal across group. The reference indicator (i.e., the item on which these equality constraints were imposed) was chosen based on an inspection of item contents to determine which were most likely to show invariance over both gender and time. For the hyperactivity/inattention factor *restless, overactive and cannot*

stay still long was used as the reference indicator; for conduct problems *generally obedient* was used; for emotional problems *often seems worried* was used; for prosociality *often volunteers to help others* was used and for peer problems '*generally liked by other children'* was used. Theta parameterization was used to facilitate testing of residual invariance. As a further model identification constraint in the configural model, item residual variances were fixed to 1 in the reference group (Millsap & Yun-Tein, 2004).

If the configural model showed acceptable fit, we proceeded to test metric invariance by imposing cross-group equality constraints on the factor loadings, that is, the loading for each item was fixed equal to the loadings for the corresponding item across males and females. There are no universally agreed on criteria for establishing invariance within categorical invariance contexts, therefore, we report both (scaled) chi-square difference tests and comparisons of comparative fit index (CFI), root mean square error of approximation (RMSEA) and standardized root mean square residual (SRMR). The disadvantage of the chi-square different tests is that they are liable to be sensitive to trivial misspecifications in the current context given the large sample size (Yuan & Chan, 2016); however, the disadvantage of RMSEA, CFI, and SRMR comparisons is that they are less well-validated for categorical indicators, with no previous study yet having evaluated their performance for the particular combination of number of factors, indicators, group-by-time point number, and sample size for ordinal indicators (Svetina et al., 2020). RMSEA, CFI, and SRMR comparisons served as our primary guide to determining invariance given the very large sample size. We used the criteria of Chen (2007), which suggests that metric invariance holds if CFI decreases by no more than .010; if RMSEA increases by no more than .015, and if SRMR increases by no more than .030. We note that other studies have suggested alternative (both more and less strict) criteria (e.g., Finch & French, 2018; Rutkowski & Svetina, 2017; Svetina & Rutkowski, 2017); however, we selected Chen's (2007) criteria because we judged them to be well-suited to the goal of identifying measurement invariance violations where they may have a nontrivial biasing effect on multigroup/longitudinal models. Chen's (2007) criteria were also used for scalar and strict invariance. If metric invariance did not hold, modification indices and expected parameter changes were used to guide the release of constraints to achieve a partially invariant model.

If a (partially) metric invariant model was achieved, we proceeded to test scalar invariance. Here we fixed all item thresholds to be equal across group and followed the same logic for determining invariance as outlined in relation to metric invariance. Chen's (2007) criteria for scalar invariance is that it holds if CFI decreases by no more than .010, if RMSEA increases by no more than .015, and SRMR increases by no more than .010 (Chen, 2007). Scalar invariance constraints were not imposed on any items that did not show metric invariance. If the initial set of scalar invariance constraints resulted in a substantial deterioration in fit according to Chen's (2007) criteria, we iteratively removed constraints, guided by modification indices and expected parameter changes in order to try and identify a partially scalar invariant model.

Finally, if a partially scalar invariant model could be achieved, we proceeded to test residual invariance. Chen's (2007) criteria are that residual invariance holds if CFI decreases by no more than .010, RMSEA decreases by no more than .015, and SRMR decreases by no more than .010 with the addition of cross-group equality constraints on residual variances.

After testing cross-group invariance, we proceeded to test longitudinal invariance, using methods based on Liu et al. (2017). Here we also used a "forwards" approach. Identification constraints in the configural model were analogous to the gender invariance model (the mean and variance of the factors at the first time point were fixed and one threshold for each item was fixed, together with loading and threshold equality constraints over time). However, it was also necessary to constrain all residual variances as well as scale factors in the configural model to achieve convergence for the baseline configural model. As such the models testing metric and scalar constraints included residual invariance constraints too so that an appropriate comparison against the configural model could be made. As a final step, we thus compared a model with metric, scalar, and residual constraints to a model with only metric and scalar constraints to assess the tenability of the residual variance constraints. These latter models were compared using a chi-square difference test and based on CFI, RMSEA, and SRMR differences. We also considered information from expected parameter changes and modification indices from the more constrained model altogether to check no substantive noninvariance had been missed.

Where adaptations to the items administered at age 3 were required to improve age-appropriateness at this earlier developmental stage, these items were paired with their corresponding item administered at later ages "as if" they were the same item, allowing their equivalence over time to be addressed as an empirical question.

Models were fit in M*plus 8.4* (Muthén & Muthén, 2015) using weighted least squares means and variances adjusted estimation. Analyses were weighted according to the longitudinal weight for Sweep 7 with stratification into the nine strata of MCS and clustering within households also accounted for by use of a sandwich estimator. The Sweep 7 longitudinal weight includes both design and nonresponse weighting and thus account for nonrandom attrition under the assumption of missingness at random (Rubin, 1976)

Table 2. Fits for Single-Group CFA Models for Males and Females.

	CFI	TLI	RMSEA	SRMR	Link to full output	CFI	TLI	RMSEA	SRMR	Link to full output
Age	Male					Female				
3	**.887**	**.872**	**.023**	**.084**	https://osf.io/96jp2/	**.849**	**.829**	**.026**	**.081**	https://osf.io/amsnb/
5	.936	.928	.023	.071	https://osf.io/phcj6/	.917	.906	.025	.073	https://osf.io/kbqg3/
7	.934	.926	.028	.061	https://osf.io/snp2j/	.923	.913	.025	.071	https://osf.io/vmfb2/
11	.905	.893	.023	.061	https://osf.io/zqdme/	.936	.927	.026	.064	https://osf.io/yfpg4/
14	.939	.931	.029	.064	https://osf.io/nw3gp/	.934	.925	.029	.074	https://osf.io/cqx32/
17	**.873**	**.857**	**.020**	**.074**	https://osf.io/mfybg/	.959	.954	.018	.069	https://osf.io/mzvta/

Note. Boldface denotes fit judged insufficient to justify inclusion in further invariance analyses. CFI = comparative fit index; TLI = Tucker–Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual

Results

Model fits for all age-by-gender subsamples are provided in Table 2. Links to the full model outputs are provided in this same table. Fit was good according to RMSEA at all ages but only poor-to-acceptable by other fit criteria. Fit was poorest at age 3 but tended to be better at older ages, except in the age 17 male group where fit was again poor in the male group. Given the poor fit in these three (time-by-gender) groups we did not include them in further examinations of cross-group or longitudinal invariance and concluded on the basis of these initial single-group CFAs that configural invariance did not hold for these groups.

Gender Invariance

Fits for the gender invariance analyses are provided in Table 3. At age 5, there were significant chi-square difference tests at all stages of invariance testing; however, CFI, RMSEA, and SRMR all either improved or declined within acceptable limits. Furthermore, modification indices and expected parameter changes did not tend to suggest any release of constraints that would lead to substantial improvements in fit. We, therefore, concluded that gender invariance up to the residual level held at age 5.

At age 7, a nonsignificant chi-square difference test, an improvement in CFI and no change in RMSEA and SRMR across the configural and metric models suggested that metric invariance held. The chi-square difference tests were significant at the scalar and residual stage; however, CFI, RMSEA, and SRMR changes were within defined limits. Indeed, CFI and RMSEA improved between the scalar and residual models. On balance we concluded that invariance held up to the residual stage across gender at age 7.

At age 11, the chi-square difference test was significant at $p < .05$ but not $p < .001$ and CFI and RMSEA improved with the addition of both metric and scalar invariance constraints (SRMR was unchanged), suggesting, on balance, that scalar gender invariance held at this age. The addition of residual constraints yielded a nonsignificant chi-square difference test, an improvement in CFI and RMSEA, and a slight deterioration in SRMR that was within predefined limits. This suggested that at age 11 gender invariance held up to the residual level.

At age 14, there was a significant chi-square difference test at the metric level accompanied by a slight improvement in CFI and RMSEA, and a slight deterioration in SRMR; however, the latter was within predefined limits. With the addition of scalar constraints, a significant chi-square difference test was accompanied by a further improvement in CFI and RMSEA and no change in SRMR, which on balance suggested that scalar invariance held. Finally, the addition of residual constraints yielded a nonsignificant chi-square difference test, improvements in CFI and RMSEA and no change in SRMR, suggesting that gender invariance at age 14 held to the residual level.

Longitudinal Invariance

Given our finding that gender invariance held up to the residual level for ages 5 up to 14 years, we proceeded to test longitudinal invariance across this age range only. Fits for each model tests are provided in Table 4. As noted above, our baseline model included residual invariance constraints but loading and threshold constraints were only the minimum required for identification. This baseline model fit well. The addition of metric constraints to this model was associated with a significant chi-square difference test; however, CFI and RMSEA both improved and SRMR decreased only slightly and well within predefined limits. The addition of scalar constraints to this model was associated with a significant chi-square difference test; however, CFI and SRMR both increased only minimally and within the predefined limits and RMSEA remained unchanged. This model now represented a fully constrained model with loadings, thresholds, and residual variances all constrained to equality over time. To explore the residual invariance assumption, this fully constrained model was compared with a model where residual variances were free to vary across time. The fully constrained model had the same CFI

Table 3. Fits for Gender Invariance Tests.

	Fit			Fit difference versus baseline			$\Delta\chi^2$ Difference test			
Model	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR	$\Delta\chi^2$	*df*	*p*	Link to full model output
Age 5										
Configural	.928	.024	.072	—	—	—	—	—	—	https://osf.io/hszgc/
Metric	.930	.023	.072	−.002	. 001	.000	46.688	20	<.001	https://osf.io/yu43a/
Scalar	.932	.023	.072	−.002	.000	.000	40.678	20	.004	https://osf.io/k9sb4/
Residual	.935	.002	.073	−.003	.021	−.001	72.826	25	<.001	https://osf.io/jt2vz/
Age 7										
Configural	.931	.026	.066	—	—	—	—	—	—	https://osf.io/29zwj/
Metric	.933	.026	.066	−.002	.000	.000	30.663	20	.0598	https://osf.io/y4xh6/
Scalar	.930	.026	.068	.003	.000	−.002	240.646	20	<.001	https://osf.io/smjxc/
Residual	.932	.025	.068	−.002	.001	.000	83.181	25	<.001	https://osf.io/5wp2e/
Age 11										
Configural	.913	.025	.063	—	—	—				https://osf.io/wfn56/
Metric	.924	.023	.063	−.009	.002	.000	36.222	20	.0145	https://osf.io/aw2zh/
Scalar	.935	.021	.063	−.011	.002	.000	35.816	20	.0162	https://osf.io/9h7qp/
Residual	.943	.019	.064	−.008	.002	−.001	37.612	25	.0504	https://osf.io/bp6dy/
Age 14										
Configural	.938	.029	.069	—	—	—				https://osf.io/v9keb/
Metric	.939	.028	.070	−.001	.001	−.001	78.444	20	<.001	https://osf.io/dr6fz/
Scalar	.940	.027	.070	−.001	.001	.000	52.655	20	<.001	https://osf.io/nt5h7/
Residual	.945	.026	.070	−.005	.001	.000	30.474	25	.2070	https://osf.io/kx27m/

Note. Metric invariance criteria were that it holds if comparative fit index (CFI) decreases by no more than .010, if root mean square error of approximation (RMSEA) increases by no more than .015, and if standardized root mean square residual (SRMR) increases by no more than .030; scalar invariance criteria were that it holds if CFI decreases by no more than .010, if RMSEA increases by no more than .015, and SRMR increases by no more than .010; residual invariance criteria were that it holds if CFI decrease by no more than .010, RMSEA decreases by no more than .015 and SRMR decreases by no more than .010.

Table 4. Longitudinal Invariance Model Fits.

Model	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR	$\Delta\chi^2$	*df*	*p*	Link to full model output
Ages 5 to 14										
Configural and residual[a]	.936	.013	.058							https://osf.io/8wy4c/
Metric and residual[a]	.940	.012	.059	−.004	.001	−.001	265.519	60	<.001	https://osf.io/3djmg/
Scalar and residual[a]	.939	.012	.060	.001	.000	−.001	1418.470	59	<.001	https://osf.io/mweu7/
Metricand scalar[a]	.939	.012	.058	—	—	—	—	—	—	https://osf.io/tv43y/
Metric, scalar and partial residual[b]	.939	.012	.059							

Note. Metric invariance criteria were that it holds if comparative fit index (CFI) decreases by no more than .010, if root mean square error of approximation (RMSEA) increases by no more than .015, and if standardized root mean square residual (SRMR) increases by no more than .030; scalar invariance criteria were that it holds if CFI decreases by no more than .010, if RMSEA increases by no more than .015, and SRMR increases by no more than .010; residual invariance criteria were that it holds if CFI decrease by no more than .010, RMSEA decreases by no more than .015 and SRMR decreases by no more than .010. *df* = degrees of freedom.
[a]In these models, residual invariance was assumed to facilitate estimation. [b]Constraints removed on the residual variance of item "complains of headaches, stomach aches, sickness" at Sweep 5 (age 11).

and RMSEA fit as the more relaxed model but a slightly higher SRMR value. The magnitude of the SRMR difference was .02, which is larger than the recommended threshold of .01 for the scalar versus residual invariance model comparison according to the criteria of (Chen, 2007). We, therefore, examined modification indices and expected parameter changes in the fully constrained model to identify possible invariance constraint misspecifications. These suggested the removal of the constraint on the residual variance of item "complains of headaches, stomach aches, sickness" at Sweep 5 (age 11). After the removal of this constraint the SRMR difference against the relevant

comparison model was within predefined limits and it was concluded that partial residual invariance held.

Discussion

In this study, we tested the longitudinal and gender measurement invariance of a popular measure of child mental health: the parent-reported SDQ (Goodman, 1997) over ages 3, 5, 7, 11, 14, and 17 years in the MCS. Though seldom-tested, longitudinal invariance is critical to many of the inferences drawn from developmental data, with metric invariance required to compare latent variances and covariances over development, scalar invariance required to compare latent means over development, for example, in the analysis of developmental trajectories, and residual invariance required to compare levels over time/group based on observed scores (Edwards & Wirth, 2012). Given that the SDQ data is commonly analyzed within MCS (e.g., Ahn et al., 2018; Carson et al., 2013; Heikkilä et al., 2011; Hesketh et al., 2016; Hope et al., 2014; Noonan et al., 2018; Patalay & Fitzsimons, 2016) and the SDQ is among the most popular omnibus measures of psychopathology in general, our study adds important evidence on the appropriateness of the SDQ as a measure of mental health across childhood and adolescence.

Our results demonstrated residual gender and longitudinal invariance for the parent-reported version of the SDQ in the MCS across ages 5 to 14 years up to the partial residual level (with only one residual invariance constraint required to be freed). Thus, it is likely to be valid to compare variances, covariances and means for each of the parent-reported SDQ constructs over this range of development using mixed gender samples or comparing genders using a latent variable model. Within a (random intercepts) cross-lagged panel model, for example, it may be of interest to evaluate whether the relations between emotional and behavioral problems remain constant over development to illuminate the evolution of their comorbidity. For example, two competing hypotheses: differentiation and dynamic mutualism have been proposed to characterize the changing relation between emotional and behavioral problems over time (Murray et al., 2016). The former proposes that the two symptom domains become less correlated over development while the latter proposes that they becoming increasingly correlated with time. Without metric invariance; however, it is not clear whether changes in their correlation merely reflect changes in their measurement over time. Equally, to identify how symptom levels change across different developmental stages, growth curve models can be used to examine the developmental trajectories of symptoms. However, changes in mean symptom levels over time are potentially confounded with changes in item thresholds over time, making scalar invariance important to demonstrate as a prerequisite to examining developmental trajectories.

Our results, however, suggested that the five-dimensional model that described the covariance in items in both males and females at ages 5 to 14 years did not well-describe their covariance at age 3 or in males at age 17. CFA models fit poorly in these groups, thus suggesting a lack of even configural invariance. Caution is needed when conducting comparisons using the SDQ involving these groups as differences could be attributable to differences in the meaning, manifestation and/or reporting of symptoms. It is possible, for example, that psychopathology is less differentiated at age 3 and thus a smaller number of dimensions is required to describe their covariance. Similarly, by age 17, parental monitoring/supervision has generally declined to a point where peer and self-reports may provide more reliable accounts of psychopathology, especially for males for whom the decline in parental monitoring and associated knowledge of their child in adolescence tends to be more pronounced (Laird, Pettit, Bates, et al., 2003; Laird, Pettit, Dodge, et al., 2003). Further research using, for example, cognitive interviewing and comparisons of parent-reports with other forms of assessment (e.g., observations by trained observers and clinical interviews) could help illuminate the sources of the difference.

Testing longitudinal measurement invariance should be a critical step in the development and validation of instruments used in developmental science (Edwards & Wirth, 2009) as it provides insights into which items may not be comparable over time and the implications of this for the interpretation of those scores at different ages. However, achieving longitudinal invariance must also be weighed against ensuring that an instrument remains appropriate at all ages and is capable of capturing age-specific manifestations in mental health issues. In the SDQ, for example, several items were adapted for administration at age 3 to better capture attention deficit hyperactivity disorder and conduct problems manifestations at this age. As a result, full longitudinal invariance is not necessarily the most appropriate goal in selecting items for a scale to be used across age. Rather, a majority of developmentally invariant items with a small number of noninvariant or age-specific items may provide a better compromise in terms of these competing consideration, given that partial invariance can often be sufficient (Pokropek et al., 2019). Despite the considerable attention paid to the validity and reliability of the SDQ (e.g., see Kersten et al., 2016, for a review), only two studies have examined its longitudinal invariance in the English language version and thus far and these did not extend beyond age 7 nor consider its interaction with gender invariance. Our study thus adds evidence for the comparability of scores from age 5 into adolescence and across males and females.

Our finding of scalar invariance of the parent-reported SDQ across ages 5 to 14 is somewhat more favorable than the findings observed for omnibus child psychosocial functioning instruments in the small number of studies that have

examined developmental invariance. For example, in a previous study of the Social Behavior Questionnaire, which shares origins with the SDQ, the anxiety subscale showed substantive departures from invariance (Murray, Obsuth, et al., 2017). Of note, a previous study using data partially overlapping with the current study suggested that the SDQ did not show developmental invariance across ages for all constructs either (Croft et al., 2015). Specifically, they found loading and threshold invariance across ages 3, 5, and 7 years for the conduct problems, prosociality, and hyperactivity/inattention factors but loading invariance only for the peer problems, and emotional problems factors. Our analysis used the same method but a slightly different sample (participants taking part in the MCS up to age 17), an increased number of waves of data, and conducted single-group CFAs for each gender/age combination and gender invariance analyses for each age prior to testing longitudinal invariance analyses. Based on our preliminary analyses, we did not include the age 3 data in our longitudinal invariance analyses because fit was poor for this group in single factor CFAs. The difference between our and this previous study illustrates the subjectivity of decision points in invariance testing and the sensitivity of findings to these decisions. Guiding these decision points and drawing optimal conclusions under conditions where different reasonable decisions lead to different findings is an area where there is a need for future research in invariance testing. This could be supported by simulation studies that cover a greater range of invariance contexts such as the current multigroup by longitudinal context (Kim & Willson, 2014) and/or the application of multiverse analysis frameworks (Steegen et al., 2016) to invariance testing. For example, future simulation studies covering the multigroup longitudinal context would be helpful to shed further light on which fit indices and thresholds are most effective for balancing Type 1 and Type 2 errors.

Limitations and Future Directions

It is important to note the limitations of the current study. First, as noted above standards used to assess invariance are subjective and depend on the level of noninvariance judged to be substantively important. In this study, we used Chen's (2007) criteria with the aim of striking a balance between detecting levels of noninvariance that, if left unmodeled, would be likely to substantively change conclusions in longitudinal models and between detecting large numbers of incidences of trivial violations. A further issue is that demonstrating measurement invariance does not guarantee that the scores from a measure have the same meaning across development, that is, that they show measurement *equivalence* over development (Meredith & Teresi, 2006). In fact, there have been examples where measurement invariance in scores has accompanied a shift in an underlying cognitive process (Widaman et al., 1992), illustrating that measurement invariance analyses may not be sensitive to important qualitative shifts over development. Other forms of evidence such as qualitative interviews (e.g., cognitive interviews; Collins, 2003) with informants rating young people at different stages of development could be used to illuminate whether and in what way informants understand and score symptoms for different ages.

Future studies should aim to examine the longitudinal invariance of other commonly used measures of child and adolescent psychosocial functioning across their recommended age ranges. Testing longitudinal measurement invariance remains relatively rare but should be a prerequisite to making strong inferences about developmental patterns from models that involve a comparison of variances, covariances, or means over time. Similarly, greater attention could be paid to developmental invariance during measure development (e.g., Sass, 2011). This would help ensure that at least a core subset of items (sufficient to achieve partial measurement invariance) show invariance over time and/or to provide insights into why particular items do not and the implications for the interpretation of scores at different ages (e.g., Edwards & Wirth, 2009). Such items could avoid referring to developmentally specific behaviors and contexts and focus on symptoms that are likely to be relevant irrespective of developmental stage. However, it is important not to select only developmentally invariant items during measure development as this would risk missing key age-specific manifestations of a construct. In some cases, the manifestations of a construct are so different across development that it is difficult to specify common items that can be used across development (Knight & Zerr, 2010). In these cases, longitudinal invariance analyses will be of limited utility.

Future studies should also examine the developmental invariance of the SDQ in other longitudinal datasets to evaluate the generalizability of the findings. Given that the SDQ is used across the world in a large number of different language versions, and in parent-, teacher, and self-report versions it would, similarly, be of considerable interest to further evaluate its invariance across other key categories in interaction with developmental invariance such as country (Ortuño-Sierra et al., 2015; Stevanovic et al., 2017) and informant (Rogge et al., 2018). One previous study suggested, for example, that measurement invariance of the SDQ did not hold across for 11/25 items across five European nations, which complicates cross-country comparisons of child development. Little is currently known the extent to which the SDQ yields comparable scores across diverse country contexts more broadly (Stevanovic et al., 2017). Future measurement invariance analyses can build further evidence surrounding the interpretation and comparability of item scores as administered in different formats and contexts. However, they can also potentially provide substantive insights into how psychopathology at different developmental stages is viewed by different observers and in different cultural contexts (Meredith & Teresi, 2006).

Conclusions

The parent-reported SDQ shows configural, metric and scalar gender and longitudinal invariance over ages 5, 7, 11, and 14 years (but not 3 and 17) across all its subscales in the MCS. This supports the use of the scale to compare variances and covariances and to examine developmental trajectories in emotional problems, conduct problems, hyperactivity/inattention, prosociality, and peer problems across childhood and adolescence.

Declaration of Conflicting Interests

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

Funding

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

ORCID iD


Aja Louise Murray https://orcid.org/0000-0002-9068-3188

References

Ahn, J. V., Sera, F., Cummins, S., & Flouri, E. (2018). Associations between objectively measured physical activity and later mental health outcomes in children: Findings from the UK Millennium Cohort Study. *Journal of Epidemiology & Community Health*, *72*(2), 94-100. https://doi.org/10.1136/jech-2017-209455

Archer, J. (2004). Sex differences in aggression in real-world settings: A meta-analytic review. *Review of General Psychology*, *8*(4), 291-322. https://doi.org/10.1037/1089-2680.8.4.291

Besemer, S., Loeber, R., Hinshaw, S. P., & Pardini, D. A. (2016). Bidirectional associations between externalizing behavior problems and maladaptive parenting within parent-son dyads across childhood. *Journal of Abnormal Child Psychology*, *44*(7), 1387-1398. https://doi.org/10.1007/s10802-015-0124-6

Bøe, T., Hysing, M., Skogen, J. C., & Breivik, K. (2016). The Strengths and Difficulties Questionnaire (SDQ): Factor structure and gender equivalence in Norwegian adolescents. *PLOS ONE*, *11*(5), e0152202. https://doi.org/10.1371/journal.pone.0152202

Booth, T., & Murray, A. L. (2018). Sex differences in personality traits. In V. Zeigler-Hill & T. K. Shackelford (Eds.), *Encyclopedia of personality and individual differences* (pp. 1-17). Springer International Publishing. https://doi.org/10.1007/978-3-319-28099-8_1265-1

Carson, C., Redshaw, M., Sacker, A., Kelly, Y., Kurinczuk, J. J., & Quigley, M. A. (2013). Effects of pregnancy planning, fertility, and assisted reproductive treatment on child behavioral problems at 5 and 7 years: Evidence from the Millennium Cohort Study. *Fertility and Sterility*, *99*(2), 456-463. https://doi.org/10.1016/j.fertnstert.2012.10.029

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

Cleverley, K., Szatmari, P., Vaillancourt, T., Boyle, M., & Lipman, E. (2012). Developmental trajectories of physical and indirect aggression from late childhood to adolescence: Sex differences and outcomes in emerging adulthood. *Journal of the American Academy of Child & Adolescent Psychiatry*, *51*(10), 1037-1051. https://doi.org/10.1016/j.jaac.2012.07.010

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research*, *12*(3), 229-238. https://doi.org/10.1023/A:1023254226592

Connelly, R., & Platt, L. (2014). Cohort profile: UK millennium Cohort study (MCS). *International Journal of Epidemiology*, *43*(6), 1719-1725. https://doi.org/10.1093/ije/dyu001

Copeland, W. E., Angold, A., Shanahan, L., & Costello, E. J. (2014). Longitudinal patterns of anxiety from childhood to adulthood: The Great Smoky Mountains Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *53*(1), 21-33. https://doi.org/10.1016/j.jaac.2013.09.017

Croft, S., Stride, C., Maughan, B., & Rowe, R. (2015). Validity of the Strengths and Difficulties Questionnaire in preschool-aged children. *Pediatrics*, *135*(5), e1210-e1219. https://doi.org/10.1542/peds.2014-2920

Davies, S. C. (2013). Our children deserve better: Prevention pays: Chief Medical Officer's annual report 2012. *Lancet*, *382*(9902), 1383-1384. https://doi.org/10.1016/S0140-6736(13)62004-8

de la Cruz, L. F., Vidal-Ribas, P., Zahreddine, N., Mathiassen, B., Brøndbo, P. H., Simonoff, E., Goodman, R., & Stringaris, A. (2018). Should clinicians split or lump psychiatric symptoms? The structure of psychopathology in two large pediatric clinical samples from England and Norway. *Child Psychiatry & Human Development*, *49*(4), 607-620. https://doi.org/10.1007/s10578-017-0777-1

Dekker, M. C., Ferdinand, R. F., Van Lang, N. D., Bongers, I. L., Van Der Ende, J., & Verhulst, F. C. (2007). Developmental trajectories of depressive symptoms from early childhood to late adolescence: Gender differences and adult outcome. *Journal of Child Psychology and Psychiatry*, *48*(7), 657-666. https://doi.org/10.1111/j.1469-7610.2007.01742.x

Edwards, M. C., & Wirth, R. J. (2009). Measurement and the study of change. *Research in Human Development*, *6*(2-3), 74-96. https://doi.org/10.1080/15427600902911163

Edwards, M. C., & Wirth, R. J. (2012). Valid measurement without factorial invariance: A longitudinal example. In J. R. Harring & G. R. Hancock (Eds.), *CILVR series on latent variable methodology: Advances in longitudinal methods in the social and behavioral sciences* (pp. 289-311). IAP Information Age Publishing.

Finch, W. H., & French, B. F. (2018). A simulation investigation of the performance of invariance assessment using equivalence testing procedures. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 673-686. https://doi.org/10.1080/10705511.2018.1431781

Gershon, J., & Gershon, J. (2002). A meta-analytic review of gender differences in ADHD. *Journal of Attention Disorders*, *5*(3), 143-154. https://doi.org/10.1177/108705470200500302

Gomez, R., & Stavropoulos, V. (2019). Parent ratings of the Strengths and Difficulties Questionnaire: What is the optimum factor model? *Assessment*, *26*(6), 1142-1153. https://doi.org/10.1177/1073191117721743

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

38(5), 581-586. https://doi.org/10.1111/j.1469-7610.1997.tb01545.x

Heikkilä, K., Sacker, A., Kelly, Y., Renfrew, M. J., & Quigley, M. A. (2011). Breast feeding and child behaviour in the Millennium Cohort Study. *Archives of Disease in Childhood*, *96*(7), 635-642. https://doi.org/10.1136/adc.2010.201970

Hesketh, K. R., Fagg, J., Muniz-Terrera, G., Bedford, H., Law, C., & Hope, S. (2016). Co-occurrence and clustering of health conditions at age 11: Cross-sectional findings from the Millennium Cohort Study. *BMJ Open*, *6*(11), e012919. https://doi.org/10.1136/bmjopen-2016-012919

Hope, S., Pearce, A., Whitehead, M., & Law, C. (2014). Family employment and child socioemotional behaviour: Longitudinal findings from the UK Millennium Cohort Study. *Journal of Epidemiology & Community Health*, *68*(10), 950-957. https://doi.org/10.1136/jech-2013-203673

Kersten, P., Czuba, K., McPherson, K., Dudley, M., Elder, H., Tauroa, R., & Vandal, A. (2016). A systematic review of evidence for the psychometric properties of the Strengths and Difficulties Questionnaire. *International Journal of Behavioral Development*, *40*(1), 64-75. https://doi.org/10.1177/0165025415570647

Kim, E. S., & Willson, V. L. (2014). Testing measurement invariance across groups in longitudinal data: Multigroup second-order latent growth model. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(4), 566-576. https://doi.org/10.1080/10705511.2014.919821

Knight, G. P., & Zerr, A. A. (2010). Informed theory and measurement equivalence in child development research. *Child Development Perspectives*, *4*(1), 25-30. https://doi.org/10.1111/j.1750-8606.2009.00112.x

Laird, R. D., Pettit, G. S., Bates, J. E., & Dodge, K. A. (2003). Parents' monitoring-relevant knowledge and adolescents' delinquent behavior: Evidence of correlated developmental changes and reciprocal influences. *Child Development*, *74*(3), 752-768. https://doi.org/10.1111/1467-8624.00566

Laird, R. D., Pettit, G. S., Dodge, K. A., & Bates, J. E. (2003). Change in parents' monitoring knowledge: Links with parenting, relationship quality, adolescent beliefs, and antisocial behavior. *Social Development*, *12*(3), 401-419. https://doi.org/10.1111/1467-9507.00240

Leopold, D. R., Christopher, M. E., Burns, G. L., Becker, S. P., Olson, R. K., & Willcutt, E. G. (2016). Attention-deficit/hyperactivity disorder and sluggish cognitive tempo throughout childhood: Temporal invariance and stability from preschool through ninth grade. *Journal of Child Psychology and Psychiatry*, *57*(9), 1066-1074. https://doi.org/10.1111/jcpp.12505

Liu, Y., Millsap, R. E., West, S. G., Tein, J.-Y., Tanaka, R., & Grimm, K. J. (2017). Testing measurement invariance in longitudinal data with ordered-categorical measures. *Psychological Methods*, *22*(3), 486-506. https://doi.org/10.1037/met0000075

Liu, Y., & West, S. G. (2018). Longitudinal measurement non-invariance with ordered-categorical indicators: How are the parameters in second-order latent linear growth models affected? *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(5), 762-777. https://doi.org/10.1080/10705511.2017.1419353

Martel, M. M. (2013). Sexual selection and sex differences in the prevalence of childhood externalizing and adolescent internalizing disorders. *Psychological Bulletin*, *139*(6), 1221-1259. https://doi.org/10.1037/a0032247

Mathyssek, C. M., Olino, T. M., Hartman, C. A., Ormel, J., Verhulst, F. C., & Van Oort, F. V. (2013). Does the Revised Child Anxiety and Depression Scale (RCADS) measure anxiety symptoms consistently across adolescence? The TRAILS study. *International Journal of Methods in Psychiatric Research*, *22*(1), 27-35. https://doi.org/10.1002/mpr.1380

McDonald, R. P. (1999). *Test theory: A unified treatment*.

Meredith, W., & Teresi, J. A. (2006). An essay on measurement and factorial invariance. *Medical Care*, *44*(11), S69-S77. https://doi.org/10.1097/01.mlr.0000245438.73837.89

Millsap, R. E. (2012). *Statistical approaches to measurement invariance*. Routledge. https://doi.org/10.4324/9780203821961

Millsap, R. E., & Yun-Tein, J. (2004). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*(3), 479-515. https://doi.org/10.1207/S15327906MBR3903_4

Molenaar, D., & Borsboom, D. (2013). The formalization of fairness: Issues in testing for measurement invariance using subtest scores. *Educational Research and Evaluation*, *19*(2-3), 223-244. https://doi.org/10.1080/13803611.2013.767628

Motl, R. W., Dishman, R. K., Birnbaum, A. S., & Lytle, L. A. (2005). Longitudinal invariance of the Center for Epidemiologic Studies-Depression Scale among girls and boys in middle school. *Educational and Psychological Measurement*, *65*(1), 90-108. https://doi.org/10.1177/0013164404266256

Murray, A. L., Booth, T., Eisner, M., Auyeung, B., Murray, G., & Ribeaud, D. (2019). Sex differences in ADHD trajectories across childhood and adolescence. *Developmental Science*, *22*(1), e12721. https://doi.org/10.1111/desc.12721

Murray, A. L., Eisner, M., Obsuth, I., & Ribeaud, D. (2017). Identifying early markers of "Late Onset" attention deficit and hyperactivity/impulsivity symptoms. *Journal of Attention Disorders*, *24*(13), 1796-1806. https://doi.org/10.1177/1087054717705202

Murray, A. L., Eisner, M., & Ribeaud, D. (2016). The development of the general factor of psychopathology "*p* factor" through childhood and adolescence. *Journal of Abnormal Child Psychology*, *44*(8), 1573-1586. https://doi.org/10.1007/s10802-016-0132-1

Murray, A. L., Eisner, M., & Ribeaud, D. (2019). Within-person analysis of developmental cascades between externalising and internalising problems. *Journal of Child Psychology and Psychiatry*, *61*(6), 681-688. https://doi.org/10.1111/jcpp.13150

Murray, A. L., Obsuth, I., Eisner, M., & Ribeaud, D. (2017). Evaluating longitudinal invariance in dimensions of mental health across adolescence: An analysis of the Social Behavior Questionnaire. *Assessment*, *26*(7), 1234-1245. https://doi.org/10.1177/1073191117721741

Muthén, L. K., & Muthén, B. (2015). *Mplus: The comprehensive modelling program for applied researchers: User's guide*. Muthén & Muthén.

Nærde, A., Ogden, T., Janson, H., & Zachrisson, H. D. (2014). Normative development of physical aggression from 8 to 26 months. *Developmental Psychology*, *50*(6), 1710-1720. https://doi.org/10.1037/a0036324

Niarchou, M., Zammit, S., & Lewis, G. (2015). The Avon Longitudinal Study of Parents and Children (ALSPAC) birth

cohort as a resource for studying psychopathology in childhood and adolescence: A summary of findings for depression and psychosis. *Social Psychiatry and Psychiatric Epidemiology*, *50*(7), 1017-1027. https://doi.org/10.1007/s00127-015-1072-8

Noonan, K., Burns, R., & Violato, M. (2018). Family income, maternal psychological distress and child socio-emotional behaviour: Longitudinal findings from the UK Millennium Cohort Study. *SSM: Population Health*, *4*(April), 280-290. https://doi.org/10.1016/j.ssmph.2018.03.002

Ortuno-Sierra, J., Chocarro, E., Fonseca-Pedrero, E., i Riba, S. S., & Muñiz, J. (2015). Evaluación de problemas emocionales y comportamentales: estructura interna del Strengths and Difficulties Questionnaire [The assessment of emotional and behavioural problems: Internal structure of the Strengths and Difficulties Questionnaire]. *International Journal of Clinical and Health Psychology*, *15*(3), 265-273. https://doi.org/10.1016/j.ijchp.2015.05.005

Ortuño-Sierra, J., Fonseca-Pedrero, E., Aritio-Solana, R., Velasco, A. M., de Luis, E. C., Schumann, G., Cattrell, A., Flor, H., Nees, F., & Banaschewski, T. (2015). New evidence of factor structure and measurement invariance of the SDQ across five European nations. *European Child & Adolescent Psychiatry*, *24*(12), 1523-1534. https://doi.org/10.1007/s00787-015-0729-x

Palmieri, P. A., & Smith, G. C. (2007). Examining the structural validity of the Strengths and Difficulties Questionnaire (SDQ) in a U.S. sample of custodial grandmothers. *Psychological Assessment*, *19*(2), 189-198. https://doi.org/10.1037/1040-3590.19.2.189

Parkes, A., Sweeting, H., & Wight, D. (2016). Early childhood precursors and school age correlates of different internalising problem trajectories among young children. *Journal of Abnormal Child Psychology*, *44*(7), 1333-1346. https://doi.org/10.1007/s10802-015-0116-6

Patalay, P., & Fitzsimons, E. (2016). Correlates of mental illness and wellbeing in children: Are they the same? Results from the UK Millennium Cohort Study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *55*(9), 771-783. https://doi.org/10.1016/j.jaac.2016.05.019

Patalay, P., Moulton, V., Goodman, A., & Ploubidis, G. B. (2017). Cross-domain symptom development typologies and their antecedents: Results from the UK millennium cohort study. *Journal of the American Academy of Child & Adolescent Psychiatry*, *56*(9), 765-776. https://doi.org/10.1016/j.jaac.2017.06.009

Pokropek, A., Davidov, E., & Schmidt, P. (2019). A Monte Carlo simulation study to assess the appropriateness of traditional and newer approaches to test for measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*(5), 724-744. https://doi.org/10.1080/10705511.2018.1561293

Rapee, R. M., Oar, E. L., Johnco, C. J., Forbes, M. K., Fardouly, J., Magson, N. R., & Richardson, C. E. (2019). Adolescent development and risk for the onset of social-emotional disorders: A review and conceptual model. *Behaviour Research and Therapy*, *123*(December), Article 103501. https://doi.org/10.1016/j.brat.2019.103501

Rogge, J., Koglin, U., & Petermann, F. (2018). Do they rate in the same way? Testing of measurement invariance across parent and teacher SDQ ratings. *European Journal of Psychological Assessment*, *34*(2), 69-78. https://doi.org/10.1027/1015-5759/a000445

Roza, S. J., Hofstra, M. B., Van Der Ende, J., & Verhulst, F. C. (2003). Stable prediction of mood and anxiety disorders based on behavioral and emotional problems in childhood: A 14-year follow-up during childhood, adolescence, and young adulthood. *American Journal of Psychiatry*, *160*(12), 2116-2121. https://doi.org/10.1176/appi.ajp.160.12.2116

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education*, *30*(1), 39-51. https://doi.org/10.1080/08957347.2016.1243540

Rutter, M., Caspi, A., & Moffitt, T. E. (2003). Using sex differences in psychopathology to study causal mechanisms: Unifying issues and research strategies. *Journal of Child Psychology and Psychiatry*, *44*(8), 1092-1115. https://doi.org/10.1111/1469-7610.00194

Salk, R. H., Petersen, J. L., Abramson, L. Y., & Hyde, J. S. (2016). The contemporary face of gender differences and similarities in depression throughout adolescence: Development and chronicity. *Journal of Affective Disorders*, *205*(November), 28-35. https://doi.org/10.1016/j.jad.2016.03.071

Sanne, B., Torsheim, T., Heiervang, E., & Stormark, K. M. (2009). The Strengths and Difficulties Questionnaire in the Bergen Child Study: A conceptually and methodically motivated structural analysis. *Psychological Assessment*, *21*(3), 352-364. https://doi.org/10.1037/a0016317

Sass, D. A. (2011). Testing measurement invariance and comparing latent factor means within a confirmatory factor analysis framework. *Journal of Psychoeducational Assessment*, *29*(4), 347-363. https://doi.org/10.1177/0734282911406661

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*(7), 643-656. https://doi.org/10.1177/0734282916655503

Speyer, L. G., Hall, H. A., Ushakova, A., Murray, A. L., Luciano, M., & Auyeung, B. (2020). Longitudinal effects of breast feeding on parent-reported child behaviour. *Archives of Disease in Childhood*, *106*(4). https://doi.org/10.1136/archdischild-2020-319038

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712. https://doi.org/10.1177/1745691616658637

Sterba, S. K., Copeland, W., Egger, H. L., Jane Costello, E., Erkanli, A., & Angold, A. (2010). Longitudinal dimensionality of adolescent psychopathology: Testing the differentiation hypothesis. *Journal of Child Psychology and Psychiatry*, *51*(8), 871-884. https://doi.org/10.1111/j.1469-7610.2010.02234.x

Stevanovic, D., Jafari, P., Knez, R., Franic, T., Atilola, O., Davidovic, N., Bagheri, Z., & Lakic, A. (2017). Can we really use available scales for child and adolescent psychopathology across cultures? A systematic review of cross-cultural measurement invariance data. *Transcultural Psychiatry*, *54*(1), 125-152. https://doi.org/10.1177/1363461516689215

Svetina, D., & Rutkowski, L. (2017). Multidimensional measurement invariance in an international context: Fit measure performance with many groups. *Journal of Cross-Cultural Psychology*, *48*(7), 991-1008. https://doi.org/10.1177/0022022117717028

Svetina, D., Rutkowski, L., & Rutkowski, D. (2020). Multiple-group invariance with categorical outcomes using updated guidelines: An illustration using M*plus* and the lavaan/semTools packages. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 111-130. https://doi.org/10.1080/10705511.2019.1602776

Tremblay, R. E. (2000). The development of aggressive behaviour during childhood: What have we learned in the past century? *International Journal of Behavioral Development*, *24*(2), 129-141. https://doi.org/10.1080/016502500383232

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

Van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*(4), 486-492. https://doi.org/10.1080/17405629.2012.686740

van Lier, P. A., Vitaro, F., Barker, E. D., Brendgen, M., Tremblay, R. E., & Boivin, M. (2012). Peer victimization, poor academic achievement, and the link between childhood externalizing and internalizing problems. *Child Development*, *83*(5), 1775-1788. https://doi.org/10.1111/j.1467-8624.2012.01802.x

Verhoeven, M., Sawyer, M. G., & Spence, S. H. (2013). The factorial invariance of the CES-D during adolescence: Are symptom profiles for depression stable across gender and time? *Journal of Adolescence*, *36*(1), 181-190. https://doi.org/10.1016/j.adolescence.2012.10.007

Wertz, J., Zavos, H., Matthews, T., Harvey, K., Hunt, A., Pariante, C. M., & Arseneault, L. (2015). Why some children with externalising problems develop internalising symptoms: Testing two pathways in a genetically sensitive cohort study. *Journal of Child Psychology and Psychiatry*, *56*(7), 738-746. https://doi.org/10.1111/jcpp.12333

Weyandt, L. L., Iwaszuk, W., Fulton, K., Ollerton, M., Beatty, N., Fouts, H., Schepman, S., & Greenlaw, C. (2003). The internal restlessness scale: Performance of college students with and without ADHD. *Journal of Learning Disabilities*, *36*(4), 382-389. https://doi.org/10.1177/00222194030360040801

Widaman, K. F., Little, T. D., Geary, D. C., & Cormier, P. (1992). Individual differences in the development of skill in mental addition: Internal and external validation of chronometric models. *Learning and Individual Differences*, *4*(3), 167-213. https://doi.org/10.1016/1041-6080(92)90002-V

Williamson, D., & Johnston, C. (2015). Gender differences in adults with attention-deficit/hyperactivity disorder: A narrative review. *Clinical Psychology Review*, *40*(August), 15-27. https://doi.org/10.1016/j.cpr.2015.05.005

Yuan, K.-H., & Chan, W. (2016). Measurement invariance via multigroup SEM: Issues and solutions with chi-square-difference tests. *Psychological Methods*, *21*(3), 405-426. https://doi.org/10.1037/met0000080