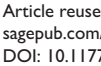
Original Research Article

The Factor Structure and Measurement Invariance of Parent-Report Strengths and Difficulties Questionnaire During a Public Health Crisis

Assessment
1–22

DOI: 10.1177/10731911251412114
journals.sagepub.com/home/asm


Simona Skripkauskaite[1], Cathy Creswell[1], Naho Morisaki[2], Aurelie Piedvache[2], and Polly Waite[1]

Abstract

The parent-report Strengths and Difficulties Questionnaire (SDQ) is a widely used child and adolescent mental health screening tool. However, challenging environments, such as public health crises, may influence the construct validity of measures. To assess this, we examine SDQ measurement invariance, internal consistency, convergent and discriminant validity, composite, test–retest, and interrater reliability across parents from the United Kingdom (n = 9,001) and Japan (n = 365). We replicate the five-factor structure, which held across children's age, gender, and between parent- and adolescent-report. We provide new evidence of SDQ invariance for special educational needs (SEN), across 6- and 1-month reporting windows, over different periods of restrictions, and between English (UK) and Japanese versions. Taken together, our findings suggest that parents interpreted the SDQ items in similar ways to pre-pandemic norms. Yet relatively low reliability of the conduct and peer relationship subscales, in particular, indicates a need for caution and scale revisions, especially when used for screening and diagnosis.

Keywords

parent SDQ, pandemic, adolescent, cross-cultural, measurement invariance

The parent-report Strengths and Difficulties Questionnaire (SDQ; R. Goodman, 1997, 2001) is a widely used screening tool for assessing mental health symptoms in 4- to 18-year-old children and has been translated into nearly 80 languages. It captures emotional symptoms, conduct problems, hyperactivity/inattention, peer relationship problems, and prosocial behaviors. Several large reviews (e.g., Kersten et al., 2016; Stone et al., 2010) have shown that the SDQ's structure is generally reliable and demonstrates measurement invariance across informants, developmental stages, and both clinical and community settings. However, evidence from cross-cultural studies is more mixed (Stevanovic et al., 2017). While adolescent self-report SDQ studies across countries often support only partial or configural invariance (Essau et al., 2012; Ortuño-Sierra et al., 2015), especially when languages or cultural contexts differ substantially (Sourander et al., 2024; Stevanovic et al., 2015), much less is known about the parent-report version. Evidence from multi-ethnic samples in Germany (Runge & Soellner, 2019) and the Netherlands (Mieloo et al., 2013) suggests partial support for the five-factor structure, albeit with some item- and subscale-level variation. This disparity is likely to reflect a combination of logistical, conceptual, and systemic factors, as parent-report data can be harder to collect cross-nationally and adolescent self-reports are often prioritized for their sensitivity to internalizing symptoms. Recent multi-country work also underscores the limited availability of parent-report measures that are meaningfully comparable across cultures (e.g., McMahon et al., 2025). Publication bias likely adds to the gap, with studies on parent-report invariance remaining unpublished when results are weak or inconclusive (c.f., Ioannidis, 2014; Putnick & Bornstein, 2016). As a result, the literature likely underrepresents variability in how parents from different cultures interpret and report child behavior. Such cross-cultural evaluations are particularly important but

[1]University of Oxford, UK
[2]National Center for Child Health and Development, Setagaya, Japan

Corresponding Author:
Simona Skripkauskaite, Department of Experimental Psychology, University of Oxford, Anna Watts Building, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK.
Email: simona.skripkauskaite@psy.ox.ac.uk

lacking for ensuring that symptom structures are interpreted consistently across linguistically and culturally distinct contexts, such as between individualistic Western countries and collectivist East Asian settings like Japan, especially when the SDQ is used in international research or global mental health monitoring.

The SDQ has been particularly instrumental in assessing child and adolescent mental health symptoms during public health crises, including natural disasters, such as earthquakes (e.g., Usami et al., 2014), bombings (e.g., Kerns et al., 2014), and the COVID-19 pandemic[1] (Bussières et al., 2021; Guzman Holst et al., 2023; Hu & Qian, 2021; Vizard et al., 2020). However, such crises may not only affect the levels of children's socio-emotional symptoms, but also how those behaviors are perceived and rated by others, potentially altering the construct validity and reliability of measurement tools like the SDQ. For instance, during public health crisis, the parent-report item "Rather solitary, tends to play alone" or the adolescent-report item "I am usually on my own" may represent limited opportunities for socializing rather than internal states such as withdrawal or loneliness. This phenomenon, known as "response shift" (Little, 2013; Oort, 2005), occurs when changes in environmental context alter the meaning of item responses, even if underlying symptom severity has not changed. Empirical work, using alternative measures, supports this concern. Olino et al. (2021) found that while measures of anxiety symptoms in adolescents and young adults remained stable from pre- to during the COVID-19 pandemic, measures of depression and intolerance of uncertainty were not psychometrically invariant over time. Yet, formal psychometric investigations of the SDQ during a public health crisis are scarce. In particular, measurement invariance testing, which is necessary to ensure that the comparisons across different groups or timepoints are meaningful and unbiased, is still rarely conducted in psychological studies more broadly (Maassen et al., 2023).

Existing Evidence of SDQ Factor Structure During a Public Health Crisis

To our knowledge, only three studies so far have evaluated the structural properties of the parent-report SDQ during a public health crisis (all within the context of COVID-19 pandemic) and none have replicated the original five-factor structure without substantial modifications. Kai Yee et al. (2022) analyzed data from Malaysian parents of 9- to 11-year-olds collected during the pandemic in 2020 (exact timing not specified). They found that the five-factor structure, identified in previous studies, only partially satisfied the criteria for acceptable model fit even after the removal of two poorly loading items ("steals from home, school, or elsewhere"; "gets along better with adults than with other children") and that internal consistency was low for the peer problem subscale. Foley et al. (2023) used data collected between April and July 2020 (i.e., the early months of the pandemic) to evaluate the cross-cultural measurement invariance of the parent-report SDQ for 3- to 8-year-old children across six countries: Australia, China, Italy, Sweden, the United Kingdom, and the United States. After the removal of three items due to low endorsement ("bullied by other children"; "fights with other children"; "steals from home or school"), their five-factor model reached a good model fit and achieved partial cross-cultural scalar invariance across countries. They found that this modified scale yielded acceptable composite reliability across subscales. More recently, Sapin et al. (2024) examined data collected later during the pandemic (Spring 2021) from parents of 3- to 17-year-old children and young people in France. They found support for the five-factor model of the SDQ when using Exploratory Structural Equation Modeling (ESEM), where cross-factor loadings of five items across conduct and hyperactivity/inattention scales were allowed. Yet, confirmatory factor analysis (CFA) models, that do not allow cross-factor loadings, showed poor model fit across five-, three-, or second-order factor structures. Their five-factor model in ESEM reached scalar invariance across range of characteristics, including parental gender, education, anxiety, depression, and psychiatric history, as well as child age and gender. It also achieved acceptable test–retest reliability across the subscales when compared with data collected in Autumn 2022. They found mostly acceptable composite reliability across subscales, although it was low for the conduct problem subscale. Together, these pandemic-based studies indicate that the parent-report SDQ may require item or model modifications to achieve acceptable fit under crisis conditions, but also raise the question of whether observed differences reflect contextual effects of the pandemic, true shifts in symptom structure, or limitations in cross-cultural measurement invariance.

Outstanding Questions

The COVID-19 pandemic presents a unique test case for evaluating the SDQ's psychometric integrity under large-scale, real-world contextual change. It raises the question of whether standard measures of mental health, such as the SDQ, remain psychometrically sound when the child's environment changes dramatically through the loss of support structures, disrupted social relationships, restricted school and leisure activities, and ongoing stress and uncertainty about the future. This is especially relevant as the SDQ is often used to track changes in mental health symptoms over time, compare different demographic groups, informants, or countries, and assess intervention outcomes. Measurement invariance provides a framework for evaluating whether such comparisons reflect true variation in

underlying constructs rather than artifacts of measurement (Van De Schoot et al., 2015; Widaman et al., 2010). Metric invariance assesses whether items relate to their latent constructs in the same way across conditions; scalar invariance evaluates whether individuals with the same underlying level of difficulties receive comparable item scores; and strict invariance tests whether item-specific measurement error varies systematically across groups or contexts. Despite its widespread use, several important gaps remain in the formal psychometric assessments of SDQ during a public health crisis.

First, it is still unclear whether the SDQ measures the same construct across time, especially during major societal disruptions such as the COVID-19 pandemic. Across different periods of public health restrictions, such as lockdowns, school closures, or relaxed distancing measures, children's experiences and social environments changed substantially. These contextual shifts may have influenced how behaviors were expressed, perceived, and reported. As a result, longitudinal studies must account not only for "true" developmental or situational change but also for potential changes in how items are interpreted over time (Little, 2013; Oort, 2005). For instance, the SDQ item "picked on or bullied by other children" is designed to capture peer victimization, which is typically observable in school or social settings. Yet, during lockdowns or remote schooling compared with the periods of no restrictions, such behavior may have been far less visible (or altogether absent), leading to underreporting or reinterpretation of the item. In such cases, observed score differences might reflect contextual reinterpretation (e.g., shifts in metric or scalar properties) rather than actual change in the underlying construct. Longitudinal invariance testing is needed to examine whether the relationships between latent constructs and the observed items used to measure them remain stable across changing conditions (Widaman et al., 2010).

Second, it is unclear how the SDQ constructs hold across different SDQ versions that have different reporting windows. The SDQ has two different formats that differ in the length of period that symptoms are reported on; the standard SDQ asks the respondent to report over the past 6 months, whereas the reference period for the follow-up version is 1 month to increase the chance of detecting change. While the two versions differ in intended temporal focus, they are designed to assess the same underlying constructs using the same items. These versions are also often used interchangeably in practice, particularly in longitudinal studies and post-intervention assessments in treatment contexts. However, variation in recall periods can influence the frequency and severity of symptoms reported (e.g., scalar non-invariance) and may threaten the comparability of measurements over time. Prior research suggests that changes in timeframe can undermine longitudinal invariance in other mental health questionnaires (e.g., PHQ-8; Moehring et al., 2021). While multiple pre-pandemic studies have demonstrated that the factorial structure of the parent-reported SDQ is longitudinally invariant across children's ages (DeVries et al., 2017; Murray et al., 2021; Sosu & Schmidt, 2017; Toseeb et al., 2022), it remains unknown whether changes in the SDQ's reference period affect temporal stability of its factorial structure.

Third, the pandemic disproportionately affected children with special educational needs (SEN), who consistently showed markedly elevated levels of emotional, conduct, and hyperactivity/inattention symptoms (Oakes et al., 2023; Waite et al., 2021). It remains unknown, however, whether the SDQ functions equivalently for children with and without SEN during times of crisis. Limited pre-pandemic evidence suggests five-factor measurement invariance for parent-reported SDQ symptoms in children with and without SEN (DeVries et al., 2018) and between community and broader clinical samples (Smits et al., 2016). However, some researchers have questioned the SDQ's structural validity for certain neurodivergent populations, such as autistic adolescents, even outside of pandemic conditions (Turcan et al., 2024). In this study, we define SEN according to parental report of formal identification under U.K. educational guidelines (Department for Education & Department of Health and Social Care, 2014), which include a range of neurodevelopmental, learning, and emotional-behavioral difficulties affecting child's learning or access to educational facilities. Understanding whether the SDQ operates consistently across these groups of children is needed for ensuring equitable assessment and valid interpretation of group differences.

Fifth, age and gender are central axes along which children's mental health symptoms often vary and not just in prevalence, but also in how behaviors are interpreted by adults. For instance, externalizing behaviors may be considered more normative at younger ages or more tolerated in boys than girls, which may influence how parents respond to SDQ items. Importantly, pre-pandemic research suggests that the parent-report SDQ demonstrated good reliability, validity, and measurement stability across age and gender groups (DeVries et al., 2017; Murray et al., 2021; Sosu & Schmidt, 2017; Toseeb et al., 2022), indicating minimal bias in typical conditions. However, the COVID-19 pandemic introduced shifts in children's routines and developmental opportunities (e.g., social play, classroom engagement), which may have altered how certain behaviors are perceived across age and gender lines. Therefore, re-establishing measurement invariance across these groups is essential to ensure that observed differences in symptom levels reflect true variation rather than shifts in how underlying levels of difficulties are interpreted (e.g., scalar non-invariance). Without this step, developmental or gender-related comparisons risk being confounded by bias in how behaviors are measured and interpreted.

Furthermore, discrepancies between parent-reported (Creswell et al., 2021) and adolescent-reported (Knowles et al., 2022) SDQ changes during the pandemic point to the need for interrater measurement invariance assessments. Such evaluations are necessary to determine whether observed differences reflect an actual mismatch of symptom perception or discrepancies in the questionnaire itself (Olino et al., 2018). Concerns have also previously been raised about the reliability of the parent-reported conduct problems and peer problems subscales (Ribeiro Santiago et al., 2022). Moreover, even prior to the pandemic, Booth et al. (2023) found that five-factor measurement invariance between parent and adolescent reports did not hold, indicating that the SDQ may function differently across informants. To date, however, the SDQ psychometric properties during the pandemic have only been investigated for parental reports, with no interrater comparisons conducted.

Finally, to ensure that psychometric properties of the parent-report SDQ are globally stable and can be used across different cultures, invariance assessments are needed (Matsumoto & Van de Vijver, 2010). While the five-factor structure of the Japanese version of the parent-reported SDQ has been previously validated within Japan (Matsuishi et al., 2008; Shibata et al., 2015), these efforts have largely focused on internal consistency and model fit within that single cultural context, without direct comparison to the English-language original. Yet, even carefully translated items can take on different meanings in different cultural contexts (Stevanovic et al., 2017). Moreover, marked cultural differences between collectivistic societies like Japan and more individualistic Western countries (e.g., the United Kingdom) may influence how behaviors such as peer conflict, social withdrawal, or emotional expression are interpreted by parents (Tashiro & Shaw, 2020). These cultural dynamics may affect both item functioning and latent construct definitions, which corresponds to potential violations of metric and scalar measurement invariance. Indeed, for adolescent self-report, prior cross-national studies have frequently found only partial or configural invariance when cultural or linguistic contexts diverge significantly (e.g., Sourander et al., 2024; Stevanovic et al., 2015). However, far fewer studies have examined whether the parent-report SDQ is similarly affected (e.g., Mieloo et al., 2013; Runge & Soellner, 2019) and none did so for Japan. Direct measurement invariance testing between the U.K. and Japanese versions offers a necessary step toward evaluating the global structural validity of the scale, particularly when comparing symptom levels or evaluating outcomes of global public health crises.

External Validity Within the SDQ Nomological Network

Beyond structural and cross-group validity, psychometric evaluations also consider reliability and external validity. In particular, it is essential to consider whether widely used measures, such as the SDQ, maintain their expected relationships with other indicators of child mental health during a public health crisis. Establishing convergent validity requires that SDQ subscales correlate strongly with measures that assess similar constructs (e.g., emotional symptoms with psychological distress), whereas discriminant validity requires that these associations are stronger than correlations with conceptually distinct constructs (Campbell & Fiske, 1959). Prior work generally shows that emotional symptoms are strongly associated with general indicators of psychological distress (e.g., Essau et al., 2012). In contrast, conduct problems and hyperactivity/inattention reflect externalizing behaviors and should demonstrate weaker associations with distress-based measures, consistent with the well-established distinction between internalizing and externalizing domains (Krueger & Markon, 2006). Prosocial Behavior is conceptually orthogonal to distress and is expected to correlate negatively with measures of psychological distress (e.g., Zhang et al., 2023). Peer Problems may fall between these domains, reflecting partial overlap with internalizing symptoms via social withdrawal or interpersonal strain (Achenbach et al., 2017). Multi-informant research further shows that ratings of the same SDQ subscale across informants (e.g., parent-adolescent emotional symptoms) generally yield stronger convergence than ratings of different subscales across reporters (e.g., Hill & Hughes, 2007). Together, these patterns outline a clear nomological network against which convergent and discriminant validity of the SDQ can be evaluated.

Present Study

The main goal of the present study was to examine the factor structure and measurement invariance of the parent-report SDQ. We also set out to assess relevant reliability and validity indicators, including those situated within the SDQ's expected nomological network. To achieve that, we have leveraged the parent- and adolescent-reported data during the COVID-19 pandemic from the Co-SPACE longitudinal survey in the United Kingdom (March 2020–July 2021) and parent-reported data from the CORONA x CODOMO study in Japan (February-March 2021). Specifically, we aimed to:

1. Replicate the five-factor structure of the parent-report SDQ scale.
2. Assess whether the measurement invariance of SDQ constructs differed between-participants across children's demographic characteristics, including age, gender, and SEN.
3. Investigate whether measurement invariance held over time across the questionnaire versions (baseline and follow-up) and across different periods of restrictions (lockdown and no restrictions).

4. Assess whether the SDQ constructs differed between raters (parents or adolescents) and across cultures (the United Kingdom or Japan).
5. Report on internal consistency, composite reliability, convergent and discriminant validity, as well as test–retest and interrater reliability across subscales.

Method

Study Design and Participants

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. The main study sample was drawn from the UK-based longitudinal, but not nationally representative, "COVID-19: Supporting Parents, Adolescents and Children during Epidemics" (Co-SPACE) study. Parents and carers (hereafter known as "parents") of children and young people aged between 4 and 16 years and their 11- to 16-year-old adolescents who lived in the United Kingdom were eligible to take part. In total, 9,208 parents took part (i.e., reached at least the questionnaire section of the survey) between April 17, 2020, and July 31, 2021. Of this group, 207 were excluded during the data cleaning as they were suspected to be "bots," were participating in an intervention trial (the "SPARKLE" trial; Kostyrka-Allchorne et al., 2021), were missing "baseline" survey data, or were missing any of the SDQ items (for information on attrition see Table S1). Thus, the final full sample consisted of 9,001 parents who took part in the Co-SPACE study at least once. This sample was further divided into five independent subsamples for the interrater, cross-cultural, longitudinal restrictions, different SDQ versions, and multi-group measurement invariance assessments (Figure 1 and Table 1). The multi-group measurement invariance subsample was then used to devise three further propensity score matched (using *MatchIt* package in R; Ho et al., 2011) subsamples for the assessment of group measurement invariance across child age, child gender, child SEN status.

In addition to the main UK-based study sample, data for cross-cultural measurement invariance was drawn from the Japan-based CORONA x CODOMO study. This involved a serial cross-sectional design online survey which collected data on children's mental health and social experiences from parents/guardians of 0- to 17-year-old children or 6- to 17-year-old children in Japan. The survey was hosted via the National Center for Child Health and Development website and ran seven times between April 30, 2020, and December 31, 2021. The current study utilizes data from 365 parents who took part in the fifth wave of the survey (February 19 to March 31, 2021) that collected parental reports of SDQ.

Ethical approvals for the studies were provided by the University of Oxford Medical Sciences Division Ethics Committee (R69060) and the institutional review board of the National Center for Child Health and Development (2020–21).

Procedure

Participants in the United Kingdom were invited to complete a monthly online survey on the Qualtrics platform. All parents provided informed online consent at the beginning of their first (i.e., baseline) survey, which they could complete at any point between April and July 2021. Additional parental informed consent, followed by informed assent from the young person, was obtained for the adolescent survey. Parents with multiple children were asked to choose one child that they would report on each time. The adolescent survey was completed immediately after the parental survey (i.e., parents were asked to pass the device to their child for them to complete their own self-report section). A link to the follow-up survey was sent via email to each parent once a month after they had completed their baseline survey. From December 2020, participants were offered the chance to win a £50 voucher in return for their participation.

Measures

As outlined earlier, the Strengths and Difficulties Questionnaire (SDQ) is a brief mental health screening questionnaire for 4- to 18-year-old children (R. Goodman, 1997, 2001). It comprises 25 items across five subscales, assessing conduct problems, hyperactivity/inattention, emotional symptoms, peer relationship problems and prosocial behavior. Items are rated on a 3-point Likert-type scale ranging from 0 ("not at all") to 2 ("certainly true"), with positively worded items in the difficulties subscales reverse-coded. Subscale scores can be obtained by summing the responses in each of the subscales (range: 0–10). The total difficulties score can then be obtained by summing scores on conduct problems, hyperactivity/inattention, emotional symptoms, and peer relationship problems (range: 0–40), with a higher score indicating greater difficulties. In contrast, for the prosocial subscale, higher scores indicate greater prosocial behavior. As is a standard requirement for the SDQ, at the first assessment (i.e., baseline) the participants are asked about symptoms over the last 6 months, but at follow-up assessments they are asked only about the preceding month. Parent, teacher, and self-reported versions of the scale exists in many different languages. The current paper focuses primarily on the parent-reported English (UK) version of the SDQ, but also the adolescent self-reported version in English and the parent-reported version in Japanese to address specific research questions.

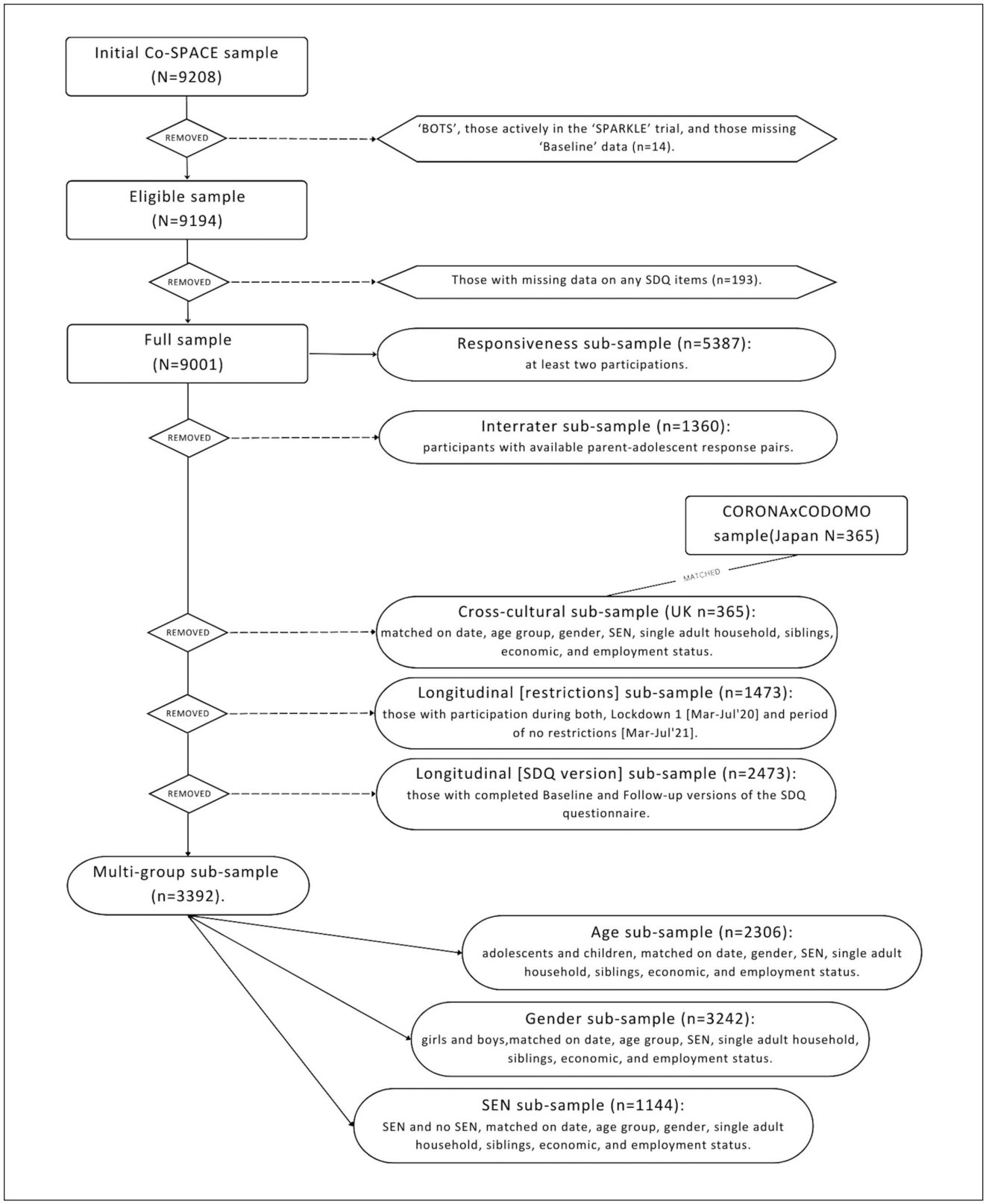


Figure 1. A Flowchart Depicting Data Processing Steps Including Cleaning and Subsampling.

Table 1. Parent-Reported Sample Characteristics per Subsampling Step.

			Cross-cultural		Longitudinal: restrictions		Longitudinal: SDQ version		
Characteristic	Full sample (baseline) (n = 9,001)	Interrater (n = 1,360)	UK (n = 365)	Japan (n = 365)	Lockdown (n = 1,473)	No restrictions (n = 1,473)	Baseline (n = 2,473)	Follow-up (n = 2,473)	Multi-group[a] (n = 3,392)
Child's age									
M (*SD*)	9.38 (3.46)	13.1 (1.71)	9.79 (3.33)	9.63 (2.86)	7.81 (2.83)	8.72 (2.83)	8.67 (3.23)	8.82 (3.26)	9.22 (3.47)
Child's age group									
Child (4–10)	5,724 (63.6%)	0 (0.0%)	232 (63.6%)	244 (66.8%)	1,271 (86.3%)	1,201 (81.5%)	1,858 (75.1%)	1,827 (73.9%)	2,239 (66.0%)
Adolescent (11–18)	3,277 (36.4%)	1,360 (100%)	133 (36.4%)	121 (33.2%)	202 (13.7%)	272 (18.5%)	615 (24.9%)	646 (26.1%)	1,153 (34.0%)
Child's gender									
Boy	4,630 (51.4%)	672 (49.4%)	163 (44.7%)	173 (47.4%)	767 (52.1%)	767 (52.1%)	1,301 (52.6%)	1,301 (52.6%)	1,751 (51.6%)
Girl	4,314 (47.9%)	667 (49.0%)	200 (54.8%)	189 (51.8%)	701 (47.6%)	699 (47.5%)	1,160 (46.9%)	1,160 (46.9%)	1,621 (47.8%)
Other/Unknown	57 (0.6%)	21 (1.5%)	2 (0.5%)	3 (0.8%)	5 (0.3%)	7 (0.5%)	12 (0.5%)	12 (0.5%)	20 (0.6%)
Special Education Needs (SEN)									
No SEN	7,512 (83.5%)	1,107 (81.4%)	319 (87.4%)	322 (88.2%)	1,243 (84.4%)	1,243 (84.4%)	2,076 (83.9%)	2,076 (83.9%)	2,820 (83.1%)
SEN	1,489 (16.5%)	253 (18.6%)	46 (12.6%)	39 (10.7%)	230 (15.6%)	230 (15.6%)	397 (16.1%)	397 (16.1%)	572 (16.9%)
Missing	0 (0%)	0 (0%)	0 (0%)	4 (1.1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Single adult household									
Non single adult	7,266 (80.7%)	1,094 (80.4%)	272 (74.5%)	302 (82.7%)	1,301 (88.3%)	1,312 (89.1%)	2,000 (80.9%)	2,005 (81.1%)	2,644 (77.9%)
Single adult	1,561 (17.3%)	251 (18.5%)	80 (21.9%)	58 (15.9%)	172 (11.7%)	161 (10.9%)	429 (17.3%)	428 (17.3%)	646 (19.0%)
Missing	174 (1.9%)	15 (1.1%)	13 (3.6%)	5 (1.4%)	0 (0%)	0 (0%)	44 (1.8%)	40 (1.6%)	102 (3.0%)
Presence of siblings									
No siblings	2,378 (26.4%)	311 (22.9%)	179 (49.0%)	137 (37.5%)	377 (25.6%)	377 (25.6%)	655 (26.5%)	655 (26.5%)	897 (26.4%)
≥ 1 sibling	6,623 (73.6%)	1,049 (77.1%)	186 (51.0%)	227 (62.2%)	1,096 (74.4%)	1,096 (74.4%)	1,818 (73.5%)	1,818 (73.5%)	2,495 (73.6%)

(continued)

Table 1. (continued)

			Cross-cultural		Longitudinal: restrictions		Longitudinal: SDQ version		
Characteristic	Full sample (baseline) (n = 9,001)	Interrater (n = 1,360)	UK (n = 365)	Japan (n = 365)	Lockdown (n = 1,473)	No restrictions (n = 1,473)	Baseline (n = 2,473)	Follow-up (n = 2,473)	Multi-group[a] (n = 3,392)
Missing	0 (0%)	0 (0%)	0 (0%)	1 (0.3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
Economic status									
≥ Average	5,907 (65.6%)	956 (70.3%)	240 (65.8%)	271 (74.2%)	1,147 (77.9%)	1,145 (77.7%)	1,545 (62.5%)	1,548 (62.6%)	2,069 (61.0%)
< Average	2,500 (27.8%)	296 (21.8%)	125 (34.2%)	91 (24.9%)	214 (14.5%)	217 (14.7%)	793 (32.1%)	787 (31.8%)	1,086 (32.0%)
Missing	594 (6.6%)	108 (7.9%)	0 (0%)	3 (0.8%)	112 (7.6%)	111 (7.5%)	135 (5.5%)	138 (5.6%)	237 (7.0%)
Parental employment status									
Employed	7,287 (81.0%)	1,148 (84.4%)	256 (70.1%)	252 (69.0%)	1,256 (85.3%)	1,268 (86.1%)	1,983 (80.2%)	1,936 (78.3%)	2,671 (78.7%)
Unemployed	1,714 (19.0%)	212 (15.6%)	109 (29.9%)	80 (21.9%)	217 (14.7%)	205 (13.9%)	490 (19.8%)	537 (21.7%)	721 (21.3%)
Missing	0 (0%)	0 (0%)	0 (0%)	33 (9.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)

[a]The multi-group sample refers to the combined age, gender, and SEN subsamples. Further breakdown for each subsample can be found in Table S2.

Adolescents in the study additionally completed the Kessler Psychological Distress Scale (K6; Kessler et al., 2002), a brief six-item self-report measure of non-specific psychological distress over the past 30 days. Items assess symptoms such as nervousness, hopelessness, restlessness, and feeling worthless, and are rated on a 5-point Likert-type scale ranging from 0 ("none of the time") to 4 ("all of the time"). Total scores range from 0 to 24, with higher scores indicating greater psychological distress. The K6 has demonstrated strong reliability and validity across adolescent populations (e.g., Mewton et al., 2016) and has also been used to assess changes in young people's mental health during the pandemic (e.g., Ferro et al., 2021).

In addition, parents reported on a number of socio-demographic variables (see Supplemental materials: Table S3). We used information on child age (primary-school-aged [4–10] or secondary-school-aged [11–17]), gender (girl or boy), and SEN status (SEN or no SEN) for multi-group comparisons. SEN status was based on parent report and reflects formal educational classifications in the United Kingdom and school-based support categories in Japan (see Supplemental materials: Table S3, for operational definitions). Other information on whether they lived in a single adult household (single adult or not single adult), presence of siblings (1 or more sibling or no siblings), household's economic status (below average or average or above), and parent's employment status (employed or unemployed) were used for sample description (Table 1) and matching (Figure 1) purposes.

Data Analysis

First, confirmatory factor analysis (CFA) on the full available sample (see Figure 1) was conducted to determine the factorial structure of the SDQ in the current sample. To achieve this, we compared the original five-factor first-order structure proposed by Goodman (1997) and broader five-factor second-order structure[2] (A. Goodman et al., 2010).

Second, the best fitting structural model was used to test measurement invariance across and within groups. Specifically, we conducted three multi-group measurement invariance assessments to determine if the measurement invariance of constructs at baseline held between-participants across children's demographic characteristics, including age (primary-school-aged [4–10] or secondary-school-aged [11–17]), gender (girl or boy), and SEN (SEN or no SEN). We then conducted two longitudinal measurement invariance tests to assess whether measurement invariance held within-participants across the questionnaire versions (baseline and follow-up), as well as at different periods of restrictions (periods of lockdowns [March–July 2020] and no restrictions [March–July 2021]). We tested longitudinal measurement invariance using a single-group correlated uniqueness CFA model. As the same participants were assessed at two timepoints in each model, the SDQ items from each timepoint were modeled as separate observed variables. Separate latent factors were also specified for each timepoint and correlations were allowed between the same latent constructs across time. To account for the repeated-measures structure, we specified correlated residuals between matching items across timepoints. Two more measurement invariance assessments were then carried out to determine if measurement invariance was present between raters (parents and adolescents) and across countries (the United Kingdom or Japan). Correlated uniqueness (i.e., residual covariances) were also included for the same items in the interrater models.

For each of these assessments, the extent of measurement invariance was evaluated iteratively as configural invariance (i.e., varying factor loadings), metric invariance (i.e., equal factor loadings), scalar invariance (i.e., equal intercepts), and strict invariance (i.e., equal residuals). As full scalar or strict invariance is rare (Putnick & Bornstein, 2016; Robitzsch & Lüdtke, 2023), the possibility of partial invariance was assessed, if needed. Decisions about which item parameters to freely estimate were guided primarily by previous studies that have adopted partial invariance approaches in SDQ measurement (e.g., Foley et al., 2023; Kai Yee et al., 2022; Ortuño-Sierra et al., 2015) as well as by inspection of modification indices and observed differences in factor loadings across groups.

Finally, we examined common validity, reliability, as well as convergent and discriminant validity coefficients for each subsample based on the best fitting factorial model. We used Cronbach's alpha (α) to assess internal consistency, omega-hierarchical (ω_h) coefficients to evaluate multidimensional composite reliability, and intraclass correlations (ICC) to assess test–retest or interrater reliability, when applicable. In line with previous systematic reviews (Bergström & Baviskar, 2021; Stone et al., 2010), internal consistency and composite reliability measures were evaluated for acceptability against the criteria of .60 to .70 as acceptable, .70 to .80 as sufficient, and above .80 as good. Subscale ICC values were evaluated against the criteria of .50 to .75 as acceptable and above .75 as good (Koo & Li, 2016). To examine external convergent and discriminant validity of parent SDQ reports, we evaluated how each parent-reported SDQ subscale related to an external measure of psychological distress (the adolescent self-reported K6) and to corresponding adolescent-reported SDQ subscales in the interrater subsample. We first computed Pearson correlations (r) between each parent-reported SDQ subscale and the K6 to assess whether subscales showed the expected association patterns with general psychological distress (e.g., strong positive correlations for emotional symptoms, weaker and/or negative correlations for externalizing and prosocial subscales). We then examined correlations between parent- and adolescent-reported SDQ subscales

to provide additional evidence of convergent validity for the SDQ itself (e.g., parent-reported—adolescent-reported emotional symptoms). Discriminant validity was evaluated by comparing these convergent associations to correlations involving theoretically distinct constructs, including (a) hetero-trait correlations between parent-reported SDQ subscales and the K6 (e.g., parent-reported conduct problems with K6) and (b) hetero-trait correlations among parent-reported SDQ subscales (e.g., parent-reported emotional symptoms with parent-reported conduct problems).

The statistical analyses were conducted in R (version 4.4.1; R Core Team, 2024). CFA models were estimated using *cfa* function in *lavaan* (version 0.6.16; Rosseel, 2012) with weighted least square mean and variance adjusted (WLSMV) estimator suitable for ordinal variables. Internal consistency and composite reliability coefficients were estimated using the *semTools* package (Jorgensen et al., 2022), intraclass correlations coefficients were produced via the *psych* package (Revelle, 2023), and differences between overlapping correlations were tested using Steiger's method for dependent correlations via the *cocor* package (Diedenhofen & Musch, 2015). The overall model fit was determined using the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and standardized root mean square residual (SRMR). Cut-off values for CFI $\geq$.950, RMSEA $<$.060, and SRMR $<$.080 were interpreted as indicating good model fit (Kline, 2015), while cut-off values for CFI $\geq$.900, RMSEA $<$.080, and SRMR $<$.080 indicated acceptable fit (Marsh et al., 2004). Due to their sensitivity to sample size, chi-square statistics were reported but not used to test model fit (Cheung & Rensvold, 2002). Instead, comparative model fit change indices were utilized. However, there is no universal consensus on optimal fit indices or cut-off values for detecting invariance, as these depend on factors such as sample size, number of comparisons, and model complexity (see Putnick & Bornstein, 2016). Following Rutkowski and Svetina (2013), we thus applied more liberal criteria (ΔCFI $\leq$.020, ΔRMSEA $>-$.030, and ΔSRMR $>-$.030 for evaluating metric invariance to reduce the likelihood of over-detecting trivial loading differences, particularly given the increased sensitivity associated with conducting multiple tests across subsamples. For scalar and strict levels, where intercept and residual constraints can have a more substantial impact on model fit and the interpretation of latent means, we adopted stricter cut-offs of ΔCFI $\leq$.010, ΔRMSEA $>-$.020, and ΔSRMR $>-$.015 (Chen, 2007).

Results

Replication of the Five-Factor Structure of the Parent-Report SDQ Scale

The original five-factor first-order structure yielded good model fit (CFI = .953, RMSEA = .057, SRMR = .058) and the broader five-factor second-order structure resulted in acceptable fit (CFI = .947, RMSEA = .060, SRMR = .061). While the more constrained second-order model did not substantially reduce model fit compared with the first-order model, convergence for it could not be achieved in multi-group assessments and thus measurement invariance is reported for first-order model only. All items in the model loaded significantly onto their expected factors (ps $<$ .05) with factor loadings ranging between .35 and .77 (Figure 2). Means and standard deviations for different subscales per subsample are available in Table S4 and Table S5.

Assessment of Whether the Measurement Invariance of SDQ Constructs Differed Between-Participants Across Children's Demographic Characteristics, Including Age, Gender, and SEN

Measurement invariance (Table 2; Models a-c) was first assessed for the U.K. parent-report SDQ asking about child's mental health symptoms over the last 6 months across four propensity-matched subsamples (Figure 1). We examined whether models could be constrained depending on child's age (primary-school-aged [4–10] or secondary-school-aged [11–17]), gender (girl or boy), and SEN status (SEN or no SEN). Acceptable model fit for the five-factor model was generally found across the multi-group invariance subsamples (Configural models in Table 2). Neither metric (ΔCFI $\leq$ .020, ΔRMSEA $>$ $-$.030, ΔSRMR $>$ $-$.030), nor scalar or strict (ΔCFI $\leq$ .010, ΔRMSEA $>$ $-$.020, and ΔSRMR $>$ $-$.015) invariance constraints resulted in significant reduction in model fit indices indicating that full psychometric measurement invariance was achieved across participants regardless of child's age or gender. Full strict invariance could not be achieved across reporting on children with and without SEN status (ΔCFI = .011) due to CFI change just exceeding the pre-set cut-off (ΔCFI $\leq$ .010). However, release of the group constraints on the residuals of a single item (item 12: "Often fights with other children or bullies them") sufficiently improved the model fit and achieved partial strict invariance.

Investigation of Whether Measurement Invariance Held Over Time Across the Questionnaire Versions (Baseline and Follow-Up) and Across Different Periods of Restrictions

We also assessed whether measurement invariance could be achieved over time within-participants in two independent subsamples (Table 2, Models d-e). Specifically, we examined whether model constraints could be imposed across the baseline and follow-up versions of the U.K. parent-reported SDQ that ask about child's mental health symptoms either over the preceding 6 months during their first participation, or the preceding month during their next participation. We then evaluated whether measurement invariance remained the same over time between first

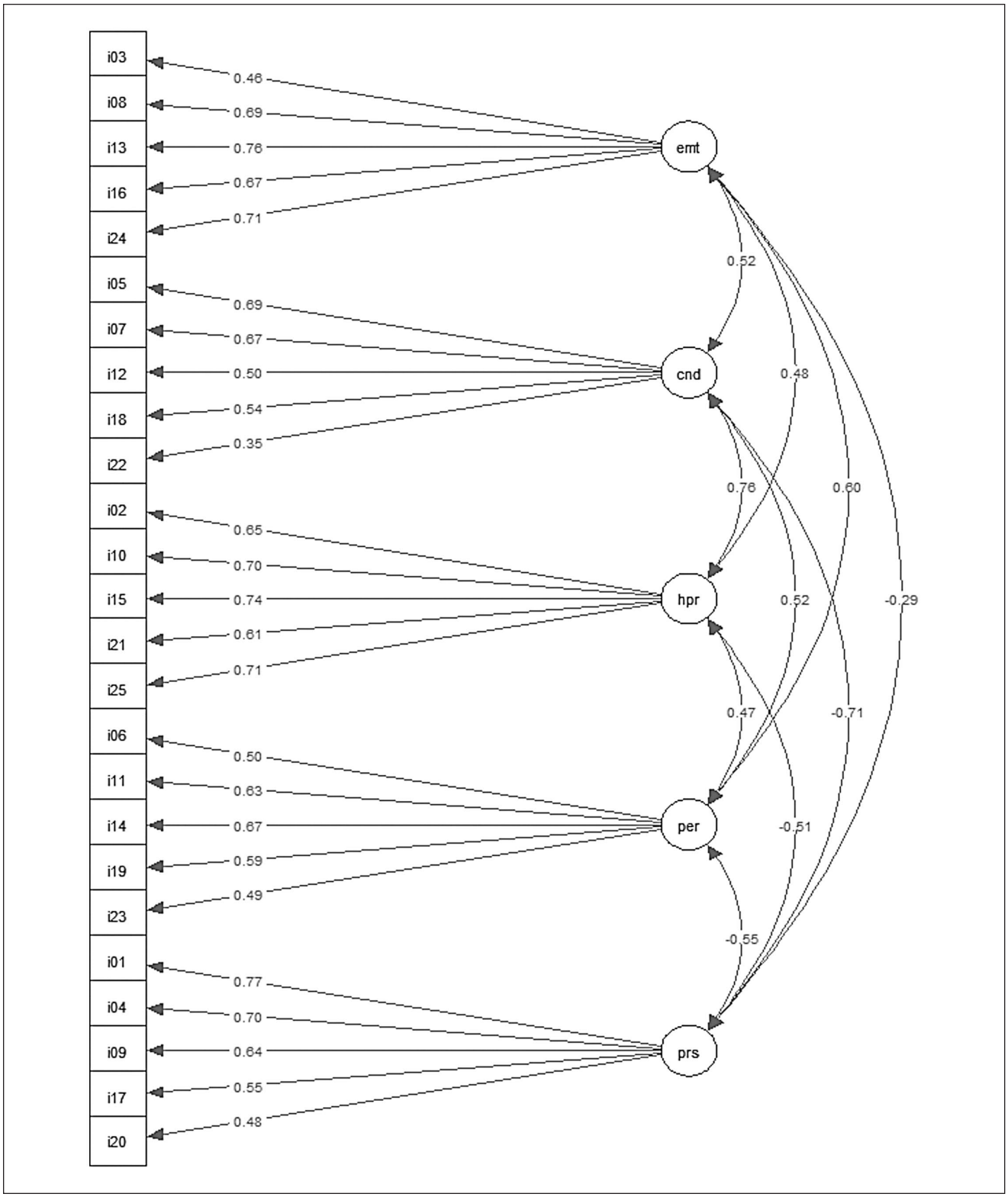


Figure 2. SDQ Factor Structure and Estimates for the Final Five-Factor First-Order Model in the Full Sample of U.K. Parents (n = 9,001).

Note. "i01"-"i25" = SDQ Item 1 to Item 25; "emt" = Emotional Symptoms; "cnd" = Conduct problems; "hpr" = Hyperactivity/ inattention; "per" = Peer relationships; "prs" = Prosocial behavior.

Table 2. Goodness-of-Fit Indices for Multigroup and Longitudinal Invariance Models.

Model	*df*	$\Delta\chi^2$	*p*	CFI	ΔCFI	RMSEA	ΔRMSEA	SRMR	ΔSRMR
a. Age invariance									
Configural	530			.949		.063		.064	
Metric	550	33.55	.029	.947	.002	.062	<.001	.065	−.001
Scalar	570	327.99	<.001	.942	.005	.063	−.001	.067	−.002
Strict	595	94.92	<.001	.941	.001	.063	.001	.068	−.001
b. Gender invariance									
Configural	530			.936		.064		.065	
Metric	550	19.91	.464	.935	.001	.063	.001	.066	−.001
Scalar	570	48.88	<.001	.935	<.001	.062	.001	.066	<.001
Strict	595	65.63	<.001	.934	.001	.061	.001	.067	−.001
c. SEN invariance									
Configural	530			.911		.074		.076	
Metric	550	56.54	<.001	.901	.009	.075	−.001	.079	−.002
Scalar	570	60.34	<.001	.899	.002	.074	.001	.080	−.001
Strict	595	221.35	<.001	.888	.011	.076	−.002	.088	−.009
Strict[a]	594	185.90	<.001	.890	.009	.076	−.001	.086	−.006
d. Long: B vs FU[b]									
Configural	1,105			.955		.053		.055	
Metric	1,130	30.02	.224	.955	<.001	.052	.001	.055	<.001
Scalar	1,155	136.11	<.001	.955	<.001	.052	<.001	.055	<.001
Strict	1,180	48.04	.004	.955	<.001	.051	.001	.055	<.001
e. Long: L vs E[b]									
Configural	1,105			.977		.043		.048	
Metric	1,130	73.65	<.001	.974	.002	.044	−.001	.050	−.002
Scalar	1,155	173.12	<.001	.974	.001	.044	<.001	.050	<.001
Strict	1,180	47.88	.004	.974	<.001	.044	<.001	.051	−.001
f. Interrater[b]									
Configural	1,105			.960		.044		.048	
Metric	1,130	140.81	<.001	.955	.006	.047	−.003	.052	−.003
Scalar	1,155	1,546.90	<.001	.942	.012	.053	−.005	.057	−.005
Scalar[c]	1,154	1,265.20	<.001	.945	.010	.052	−.004	.056	−.004
Strict[c]	1,178	362.73	<.001	.941	.003	.053	−.001	.060	−.004
g. Cross-cultural									
Configural	530			.905		.073		.081	
Metric	550	54.31	<.001	.894	.011	.076	−.003	.086	−.005
Scalar	570	224.29	<.001	.880	.014	.079	−.003	.090	−.004
Scalar[d]	569	174.37	<.001	.884	.010	.078	.002	.089	.003
Strict[d]	594	155.48	<.001	.873	.010	.080	−.002	.094	.005

Note. Metric invariance is accepted if ΔCFI ≤ .020, ΔRMSEA > −.030, and ΔSRMR > −.030, while scalar and strict invariance are accepted if ΔCFI ≤ .010, ΔRMSEA > −.020, and ΔSRMR > −.015.

[a]Item 12 residuals vary between raters. [b] Items correlated over time or over raters. [c] Item 2 intercepts vary between raters. [d] Item 21 intercepts vary between countries.

national lockdown [March–July '20] and a year later when restrictions were fully released [March–July '21]). Good model fit for the five-factor structure was found in configural models for both longitudinal subsamples. Neither metric, scalar, nor strict invariance constraints resulted in a significant reduction in model fit indices, indicating that full psychometric measurement invariance was also achieved within-participants over time regardless of SDQ version or period of restrictions.

Assessment of Whether the SDQ Constructs Differed Between Raters (Parents or Adolescents) and Across Cultures (the United Kingdom or Japan)

Two more subsamples were utilized to examine whether interrater and cross-cultural invariance could be achieved (Table 2, Models f-g). For the interrater analysis, measurement invariance was assessed for all available

parent-adolescent report pairs. A cross-cultural comparison was conducted by combining parent-report SDQ data from Japan and propensity matched parent-reports from the United Kingdom. Good configural model fit was achieved for the interrater model and acceptable configural fit was achieved across cultures. Metric invariance constraints did not result in significantly poorer model fit in either of the subsamples and, thus, could be imposed. However, full scalar and strict invariance could not be achieved for either parent-adolescent pairs (ΔCFI = .012) or across the United Kingdom and Japan (ΔCFI = .014) due to CFI change exceeding the pre-set cut-off (ΔCFI $\leq$.010). Instead, minor model re-specifications were implemented based on modification indices to achieve partial scalar and strict invariance. In the interrater subsample, the intercepts and residuals between scores of the item 2 were allowed to vary between raters (parent version: "Restless, overactive, cannot stay still for long"; adolescent version: "I am restless, I cannot stay still for long"). In the cross-cultural subsample, the intercepts and residuals between scores of the item 21 ("Thinks things out before acting") were allowed to vary between participants in the United Kingdom and Japan.

Report on Internal Consistency, Composite Reliability, Convergent and Discriminant Validity, as Well as Test–Retest and Interrater Reliability Across Subscales

Internal consistency and composite reliability of the SDQ was assessed for the U.K. parent reports across the full sample and multi-group subsamples, for adolescent self-report, and for Japanese parent reports (Table 3). In addition, test–retest reliability was examined in longitudinal subsamples and interrater reliability was estimated for parent-adolescent dyads. Convergent and discriminant of parent-reported SDQ was examined via correlations with adolescent self-reported K6 scale and adolescent-reported SDQ in an interrater sample.

Parent-Report in the U.K. Sample. Overall, Cronbach's alpha (α) indicated acceptable to good internal consistency across subscales in the full sample (α: .68–.80). Internal consistency was relatively similar across age (Primary α: .68–.81; Secondary α: .70–.82), gender (Girls α: .67–.80; Boys α: .67–.80), and SEN (Yes α: .66–.78; No α: .61–.78) subsamples. Omega-hierarchical indicated that composite reliability was sufficient to good across subscales in the full sample (ω_h: .70–.82). It was acceptable to good across age groups (Primary ω_h: .66–.81; Secondary ω_h: .65–.82), for boys (ω_h: .67–.82), and for children with SEN (ω_h: .66–.81), but slightly lower for girls (ω_h: .64–.79). For children without SEN, most subscales also reached acceptable to sufficient reliability (ω_h: .68–.77), but composite reliability for the peer relationship subscale was poor (ω_h = .51).

Test–retest reliability of hyperactivity/inattention subscale was good across different periods of restriction (ICC = .75) and between different reporting windows of the SDQ (ICC = .76). Peer relationship subscale test–retest reliability was acceptable across restriction (ICC = .74) and good between SDQ versions (ICC = .78). Conduct problems (ICC: .71–.72), emotional symptoms (ICC: .67–.74), and prosocial behaviors (ICC: .71–.74) subscales all showed acceptable test-rest reliability in both longitudinal samples, respectively.

Patterns of correlations supported the external convergent and discriminant validity of the parent-reported SDQ (Table 4). As expected, parent-reported emotional symptoms showed the strongest correlation with adolescent-reported psychological distress on the K6 (r = .57), whereas K6 correlations with conduct problems, hyperactivity/inattention, peer relationships, and prosocial behavior were weaker and in the theoretically expected directions. This was confirmed by Steiger's overlapping-correlation tests (zs = 8.79–20.92, ps < .001; see Supplemental materials: Table S6). Parent–adolescent agreement also demonstrated strong mono-trait convergence (rs = .54–.67), and these correlations consistently exceeded hetero-trait cross-informant associations (zs = 8.09–16.59, ps < .001). Together, the correlation matrix (Table 4) and comparison tests (Supplemental materials: Table S6) suggest that parent-reported SDQ subscales showed stronger relationships with conceptually aligned constructs, such as adolescent distress and matching adolescent SDQ subscales, than with unrelated traits.

Adolescent-Report in the U.K. Sample. Cronbach's alpha (α) indicated acceptable to sufficient internal consistency across all adolescent self-reported subscales (α: .62–.78). Composite reliability was good for hyperactivity/inattention (ω_h = .84) and emotional symptoms (ω_h = .83) subscales, sufficient for peer relationship (ω_h = .71) and prosocial behavior (ω_h = .79) subscales, and acceptable for the conduct problem (ω_h = .65) subscale. Interrater reliability between parents and adolescents was acceptable (ICC = .65).

Parent-Report in the Japanese Sample. Cronbach's alpha (α) also indicated acceptable to sufficient internal consistency across all subscales (α: .60–.79) in the Japanese version of the SDQ. Similarly, omega-hierarchical indicated acceptable to sufficient composite reliability across subscales (ω_h: .66–.78).

Discussion

The main aim of the current study was to evaluate the factor structure and measurement invariance of the parent-report SDQ during a global public health crisis, the COVID-19 pandemic. We replicated the original five-factor structure, which fully held across children's age and gender groups and was partially achieved between groups of children with

Table 3. Internal Consistency (α) and Composite Reliability (ω_h) of SDQ Five-Factor Model per Reporter and Child Characteristics.

	Hyperactivity/ inattention		Conduct problems		Emotional symptoms		Peer relationships		Prosocial behavior	
Sample	α	ω_h	α	ω_h	α	ω_h	α	ω_h	α	ω_h
Parent reports UK										
Full sample										
	.80	.82	.70	.70	.79	.79	.68	.70	.77	.75
Age										
Primary	.81	.81	.70	.66	.78	.81	.68	.73	.77	.76
Secondary	.80	.82	.70	.71	.82	.80	.72	.65	.77	.75
Gender										
Boys	.80	.82	.68	.69	.76	.79	.67	.67	.76	.75
Girls	.78	.78	.70	.68	.80	.79	.67	.64	.75	.73
SEN										
No	.78	.77	.70	.70	.78	.74	.61	.51	.73	.68
Yes	.76	.79	.73	.66	.76	.81	.66	.74	.78	.78
Adolescent reports UK										
	.78	.84	.62	.65	.78	.83	.65	.72	.69	.79
Parent reports Japan										
	.77	.77	.60	.66	.79	.78	.61	.66	.68	.76

or without SEN. We were able to establish longitudinal measurement invariance across parent-reported SDQ standard baseline and follow-up versions and over different periods of restrictions, as well as partial measurement invariance between parent- and adolescent-report and English (UK) and Japanese versions of the SDQ. The five SDQ subscales showed acceptable to good internal consistency and composite reliability across subsamples, although composite reliability for the peer relationship subscale in the sample of children without special educational needs was poor. The subscales also yielded, at least, acceptable test–retest reliability across different SDQ reporting windows and different periods of restrictions, as well as acceptable convergent and discriminant validity and interrater reliability between parents and adolescents, including the expected pattern of strong mono-trait associations and weaker hetero-trait relationships within the SDQ's nomological network.

Unlike the previous studies that have investigated the SDQ factorial structure during the pandemic, our findings replicated the original five-factor structure without requiring any item removals or model modifications. Both the original (R. Goodman, 1997) and broader second-order structure (A. Goodman et al., 2010) yielded a good model fit in the current study indicating that both structures could be applied to the data. However, the measurement invariance of the latter structure could not be assessed due to non-convergence issues, and it is therefore possible that this structure could have varied over time, across groups, reporters, or countries. In contrast, the five-factor structure of the SDQ showed measurement invariance across age and gender groups. This aligns with pre-pandemic research suggesting that the parent-report SDQ is a reliable, valid, and stable measure of emotional and behavioral symptoms in children (4- to 10-year-olds) and adolescents (11- to 16-year-olds), as well as across girls and boys (DeVries et al., 2017; Murray et al., 2021; Sosu & Schmidt, 2017; Toseeb et al., 2022). Taken together, these findings indicate that parents interpreted SDQ items in broadly the same way during the pandemic as in typical conditions, in contrast to earlier pandemic studies that reported weaker fit or required structural modifications (e.g., Kai Yee et al., 2022; Sapin et al., 2024).

Differences between the present findings and other pandemic studies may have occurred due to variations in language versions, cultural contexts, or modeling approaches rather than genuine instability in SDQ constructs. For instance, pandemic studies using the French (Sapin et al., 2024) and Malaysian (Kai Yee et al., 2022) parent-report versions of the SDQ encountered model-fit challenges or required item-level modifications. This aligns with pre-pandemic findings from adolescent-report studies across European (Essau et al., 2012; Ortuño-Sierra et al., 2015) and Asian (Sourander et al., 2024) countries showing that SDQ item functioning can vary across particular languages and cultural settings. Yet, these variations are difficult to distinguish from potential pandemic related effects, given that Foley et al. (2023) was the only study that included the original English version of the SDQ in their cross-country comparisons but did not report model fit across countries before they modified the scale for acceptable fit. These considerations underscore the

Table 4. Multitrait-Multimethod Correlations Among Parent-Reported SDQ, Adolescent-Reported SDQ, and Adolescent-Reported K6.

Method	Parent SDQ					Adolescent SDQ				
Trait	Hyperactivity/ inattention	Conduct problems	Emotional symptoms	Peer relationships	Prosocial behavior	Hyperactivity/ inattention	Conduct problems	Emotional symptoms	Peer relationships	Prosocial behavior
Parent SDQ										
Hyperactivity/ inattention	—									
Conduct problems	*.57*	—								
Emotional symptoms	*.47*	*.40*	—							
Peer relationships	*.38*	*.40*	*.48*	—						
Prosocial behavior	*−.39*	*−.52*	*−.26*	*−.42*	—					
Adolescent SDQ										
Hyperactivity/ inattention	**.63**	.41	.35	.28	−.27	—				
Conduct problems	.44	**.62**	.25	.27	−.37	*.53*	—			
Emotional symptoms	.30	.24	**.67**	.37	−.14	*.40*	*.28*	—		
Peer relationships	.30	.29	.41	**.65**	−.27	*.32*	*.35*	*.47*	—	
Prosocial behavior	−.23	−.33	−.18	−.23	**.54**	*−.29*	*−.35*	*−.08*	*−.40*	—
Adolescent K6										
Psychological distress	.34	.31	**.57**	.37	−.23	.39	.39	**.69**	.50	−.07

Note. Values in bold represent convergent (mono-trait/hetero-method) correlations. Italicized correlations indicate strong discriminant validity coefficients (hetero-trait/mono-method). All other values denote weak discriminant validity coefficients (hetero-trait/hetero-method).

need for careful attention to both linguistic and contextual influences when interpreting cross-study variability in SDQ structure during the pandemic.

Our findings also indicate that the parent-reported SDQ is as suitable for children with SEN as it is for children without SEN, showing no strong evidence that its psychometric performance is poorer in one group than the other. This conclusion is supported by partial strict invariance and acceptable internal consistency and composite reliability across SEN status groups, in line with the limited previous research showing measurement invariance across these groups for the self-reported SDQ in German young people aged 11 to 12 years old (DeVries et al., 2018). While our SEN group included children formally identified under educational guidance in the U.K. sample and within school-based support categories in Japan, it is important to acknowledge that this is a heterogenous group and as such, we should remain cautious about the applicability of these findings to specific neurodivergent populations. Notably, Turcan et al. (2024) found poor factorial fit for parent-reported SDQ for autistic English adolescents; however, consistent with our findings, their supplementary analyses showed improved and adequate model fit when the SDQ 3-point Likert-type response scale was treated as categorical rather than continuous (i.e., using WLSMV estimator). This suggests that the SDQ structure may remain valid for some neurodivergent groups under appropriately specified models, though factorial validity may vary across specific diagnoses.

While our current findings indicate that the factor structure of the parent-reported SDQ remained consistent across all of our measurement invariance assessments, the reliability of the subscales was generally suboptimal. This implies that subscale items might not always reflect the same construct. The conduct and peer relationship subscales, in particular, demonstrated only acceptable internal consistency and composite reliability results. Consistent with previous research indicating problems with the conduct subscale (Foley et al., 2023; Kai Yee et al., 2022; Sapin et al., 2024), item 22 ("steals from home or school") had the lowest factor loading and lowest endorsement (7.9%) in our study. It is possible that this reflects COVID-19-related restrictions on face-to-face interactions, which may have limited children's opportunities or motivation to engage in and, in turn, parents' opportunities to observe particular behaviors. However, it is important to note that similar subscale-level validity and reliability scores have commonly been found in pre-pandemic SDQ assessments (see Stone et al., 2010) including its original psychometric evaluation (R. Goodman, 2001). Indeed, when assessed, most studies evaluating SDQ psychometric properties have considered internal consistency and/or composite reliability of .60 as acceptable (Kai Yee et al., 2022; Matsuishi et al., 2008; Stone et al., 2015; Woerner et al., 2004). Yet, while values as low as .70 (Nunnally, 1978) or .60 (Hair et al., 2011) may be acceptable for exploratory research during scale development, higher reliability (e.g., .80 for basic research and .90 applied settings) is recommended (Lance et al., 2006). This is especially important given the widespread use of the SDQ. Due to its brevity, availability of norms, and accessibility, it is used not only in research, but also in clinical and social care settings. Therefore, although our findings support the use of the parent-report SDQ during the COVID-19 pandemic, including with children with SEN and in Japanese contexts, they also align with the recent calls to revise the SDQ and to use it cautiously in clinical practice, especially for screening and diagnostic purposes (Kankaanpää et al., 2023; Turcan et al., 2024).

Our findings also provide new evidence showing measurement invariance longitudinally across the standard baseline and follow-up versions of the parent-reported SDQ and across the different stages of the pandemic-related restrictions, as well as across parent and adolescent informants. Multiple pre-pandemic studies to date have investigated and confirmed that factorial structure of the parent-reported SDQ is time invariant across children's ages (DeVries et al., 2017; Murray et al., 2021; Sosu & Schmidt, 2017; Toseeb et al., 2022). Yet whether the reporting window of the SDQ that is used has potential effects on the factorial structure of the scale has not been addressed. Current findings of longitudinal strict invariance and acceptable test–retest reliability suggest that factor structure, intercepts, loadings, and residuals were invariant across the baseline and follow-up versions of the SDQ despite their differences in the length of reporting period. Similarly, strict invariance and acceptable test–retest reliability was also found between parental reports during the first national lockdown (March–July '20) and a year later when restrictions were fully released (March–July '21). Partial interrater invariance was also found between parent and adolescent informants suggesting that SDQ constructs were meaningfully consistent between parental reports and adolescent reports during the pandemic. This suggests that previously reported variations in SDQ symptom levels over the pandemic (Creswell et al., 2021) are likely to reflect a true change in parental perception of their children's mental health symptoms rather than changes in construct perception.

We also provide novel insights into global stability of the parent-reported SDQ as our findings established partial measurement invariance across the English (UK) and Japanese versions. The SDQ has been originally developed in the United Kingdom and translated into nearly 80 languages including Japanese. This version has been validated within Japan and has been found to have a five-factor structure as expected (Matsuishi et al., 2008; Shibata et al.,

2015). Yet, any group comparisons based on measures that were validated in different cultural contexts separately may yield spurious differences that occur due to differences in factor intercepts or loadings rather than observed means (Chen, 2008). "Japanese collectivism" has been suggested to underlie the markedly different experience of the COVID-19 pandemic in Japan in comparison to more individualistic Western countries, such as the United Kingdom (Tashiro & Shaw, 2020), which could also lead to different perceptions of child mental health symptoms between countries. Yet, our findings of partial strict invariance are reassuring in that regard by showing that sufficient measurement invariance is present to allow for cross-cultural comparisons between the United Kingdom and Japan.

Limitations and Constraints on Generality

The generalizability of the current findings should be interpreted in light of several limitations. First, as the Co-SPACE project was started in response to the COVID-19 pandemic, pre-pandemic SDQ psychometric properties could not be examined in the current sample. However, we showed strict longitudinal measurement invariance between parental reports during the first national lockdown and when restrictions were fully released in 2021. Furthermore, the current findings of five-factor structure and measurement invariance between reporters and across child age, gender, and SEN groups align with pre-pandemic research. However, it is still possible that more enduring shifts in parental expectations, children's social development, or the expression of symptoms could have taken place from pre- to during- and post-pandemic that could affect SDQ construct perception but would have not been captured in the current analysis. For example, behaviors such as social withdrawal may have become more normalized due to sustained remote learning, social isolation, and altered routines. Post-pandemic, parents might report these behaviors as less concerning than they would have prior to 2020. Second, the current study only investigated parent-report SDQ measurement invariance for children with and without SEN as two groups. Given that Turcan et al. (2024) found poor factorial fit for SDQ in autistic adolescents, future studies should further examine whether and how the SDQ factorial structure may vary across different types and co-occurrences of special educational needs. Third, the primary study sample from the United Kingdom was based on the convenience sampling resulting in WEIRD (i.e., western, educated, industrialized, rich, and democratic) bias with over-representation of middle and high-income parents from White British backgrounds. A relatively small sample of participants was available to assess measurement invariance of the Japanese SDQ version. Acceptable partial invariance was achieved between the United Kingdom and Japan subsamples that were propensity matched on date of participation and demographic information, but nether multi-group nor longitudinal or interrater invariance could be assessed in the Japanese sample separately. Furthermore, SEN status was based on parent report, reflecting formal identification under U.K. educational guidance in the U.K. sample, and school-based support categories in Japan. Therefore, further research is needed to establish whether current findings of the measurement invariance across age, gender, SEN groups, parent and adolescent raters, different versions of the SDQ and different stages of the pandemic based on the U.K. sample would fully apply to the Japanese version of parent-reported SDQ. Moreover, criterion validity could not be assessed in the current study due to the lack of diagnostic or clinical outcome data, largely stemming from the constraints of online data collection during the COVID-19 pandemic. This limits our ability to evaluate how well SDQ scores predict clinically significant outcomes. Finally, teacher reports were not collected and assessed in the current study despite pre-pandemic research that indicates high validity and reliability of teacher reported SDQ (A. Goodman et al., 2010). Given the increased frequency of the online learning and common changes in teacher-to-class allocation during the pandemic, it is likely that the psychometric properties of the teacher-report SDQ version may have been affected more than the parent-report, but we were not able to explore this.

Conclusion

Taken together, our results suggest that the parent-report SDQ is generally a reliable and structurally valid measure for capturing changes in child and adolescent mental health during public health crises, such as the COVID-19 pandemic, although it is not without limitations. This study replicated the original five-factor structure of the parent-reported SDQ during the COVID-19 pandemic without any modifications and full longitudinal measurement invariance was established for the data collected during the height of restrictions and when restrictions were lifted. This suggests that any the public health crisis-related construct changes to this version of the SDQ, if present, were not substantial enough to undermine the factorial structure in this study. Our findings further suggest that parents interpret the SDQ items in similar ways to pre-pandemic norms irrespective of their child's age, gender, and SEN status, 1- or 6-month reporting period, or English and Japanese language versions of the questionnaire. However, we note that the reliability of the parent-reported SDQ, in particular conduct and peer relationship, was quite low across our samples. Low reliability of these subscales is in line with previous pre-pandemic estimates, but it indicates a need for caution and scale revisions, especially when used in clinical practice for screening and diagnosis.

Acknowledgments

The authors would like to thank all the parents/carers and young people for taking part in the Co-SPACE and CORONA x CODOMO studies. The authors would also like to thank all the members of the Co-SPACE research team for their invaluable input: Lowrie Hilladakis (née Burgess), Ning Ding, Amy McCall, Martha Oakes, Samantha Pearcey, Jasmine Raw, Olly Robertson, and Adrienne Shum.

Data Availability Statement

The Co-SPACE data are partially (parent reports only) available under safeguarded access via the UK Data Service at http://doi.org/10.5255/UKDA-SN-8900-1, reference number SN 8900. The research protocols containing further procedural information are available via the Open Science Framework (OSF; Co-SPACE: https://osf.io/y8ejg). The CORONA x CODOMO study data is not publicly available due to privacy restrictions.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the UKRI/ESRC (ES/W011972/1 and ES/V004034/1), JSPS (JPJSJRP20211709 and JPMJSC21U6), JST (JPMJJR2008), and the Westminster Foundation. C.C. and P.W. receive funding from the National Institute for Health and Care Research (NIHR) Oxford and Thames Valley Applied Research Collaboration and the NIHR Oxford Health Biomedical Research Centre. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Ethical Considerations

Ethical approvals for the studies were provided by the University of Oxford Medical Sciences Division Ethics Committee (R69060) and the institutional review board of the National Center for Child Health and Development (2020–21).

Consent to Participate

All parents provided informed online consent for the parent-reported data. Additional parental informed consent, followed by informed assent from the young person, was obtained for the adolescent survey.

Consent for Publication

Not applicable.

ORCID iDs

Simona Skripkauskaite https://orcid.org/0000-0002-7501-4111
Cathy Creswell https://orcid.org/0000-0003-1889-0956
Aurelie Piedvache https://orcid.org/0000-0002-2737-0461
Polly Waite https://orcid.org/0000-0002-1967-8028

Supplemental Material

Supplemental material for this article is available online.

Notes

1. The World Health Organization declared COVID-19 a Public Health Emergency of International Concern (PHEIC) on January 30, 2020, and characterized it as a global pandemic on March 11, 2020, with the emergency phase ending on May 5, 2023 (Burki, 2023).
2. A five-factor second-order structure of the SDQ further extends the original five-factor structure and assumes an additional second-order level of latent factors representing two further SDQ subscales. In this model, hyperactivity/ inattention and conduct factors further load on an externalizing second-order factor and emotional and peer relationship first-order factors load onto an internalizing second-order factor, while the prosocial behavior subscale remains as a first-order factor.

References

Achenbach, T. M., Ivanova, M. Y., & Rescorla, L. A. (2017). Empirically based assessment and taxonomy of psychopathology for ages 1½–90+ years: Developmental, multi-informant, and multicultural findings. *Comprehensive Psychiatry*, *79*, 4–18. https://doi.org/10.1016/J.COMPPSYCH.2017.03.006

Bergström, M., & Baviskar, S. (2021). A systematic review of some reliability and validity issues regarding the Strengths and Difficulties Questionnaire focusing on its use in out-of-home care. *Journal of Evidence-Based Social Work*, *18*(1), 1–32. https://doi.org/10.1080/26408066.2020.1788477

Booth, C., Moreno-Agostino, D., & Fitzsimons, E. (2023). Parent-adolescent informant discrepancy on the Strengths and Difficulties Questionnaire in the UK Millennium Cohort Study. *Child and Adolescent Psychiatry and Mental Health*, *17*(1), 1–13. https://doi.org/10.1186/S13034-023-00605-Y

Burki, T. (2023). WHO ends the COVID-19 public health emergency. The Lancet. *Respiratory Medicine*, *11*(7), 588. https://doi.org/10.1016/S2213-2600(23)00217-5

Bussières, E. L., Malboeuf-Hurtubise, C., Meilleur, A., Mastine, T., Hérault, E., Chadi, N., Montreuil, M., Généreux, M., Camden, C., Roberge, P., Lane, J., Jasmin, E., Kalubi, J. C., Bussiéres, E. L., Hurtubise, K., Bach, G., Chrysagis, M., Turner, M. P., Gauvin, C., & Hérault, E. (2021). Consequences of the COVID-19 pandemic on children's mental health: A meta-analysis. *Frontiers in Psychiatry*, *12*, 691659. https://doi.org/10.3389/FPSYT.2021.691659

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/H0046016

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

Chen, F. F. (2008). What happens if we compare chopsticks with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social*

Psychology, *95*(5), 1005–1018. https://doi.org/10.1037/A0013193

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Creswell, C., Shum, A., Pearcey, S., Skripkauskaite, S., Patalay, P., & Waite, P. (2021). Young people's mental health during the COVID-19 pandemic. *The Lancet Child & Adolescent Health*, *5*(8), 535–537. https://doi.org/10.1016/S2352-4642(21)00177-2

Department for Education, & Department of Health and Social Care. (2014, May 11). SEND code of practice: 0 to 25 years. GOV.UK. https://www.gov.uk/government/publications/send-code-of-practice-0-to-25

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. https://doi.org/10.1016/J.PAID.2017.06.026

DeVries, J. M., Voß, S., & Gebhardt, M. (2018). Do learners with special education needs really feel included? Evidence from the Perception of Inclusion Questionnaire and Strengths and Difficulties Questionnaire. *Research in Developmental Disabilities*, *83*, 28–36. https://doi.org/10.1016/J.RIDD.2018.07.007

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLOS ONE*, *10*(4), e0121945. https://doi.org/10.1371/JOURNAL.PONE.0121945

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

Ferro, M. A., Meyer, S. B., Yessis, J., Reaume, S. V., Lipman, E., & Gorter, J. W. (2021). COVID-19-related psychological and psychosocial distress among parents and youth with physical illness: A longitudinal study. *Frontiers in Psychiatry*, *12*, 761968. https://doi.org/10.3389/FPSYT.2021.761968

Foley, S., Ronchi, L., Lecce, S., Feng, X., Chan, M. H. M., & Hughes, C. (2023). Cross-cultural equivalence of parental ratings of child difficulties during the pandemic: Findings from a six-site study. *International Journal of Methods in Psychiatric Research*, *32*(1), e1933. https://doi.org/10.1002/MPR.1933

Goodman, A., Lamping, D. L., & Ploubidis, G. B. (2010). When to use broader internalising and externalising subscales instead of the hypothesised five subscales on the Strengths and Difficulties Questionnaire (SDQ): Data from British parents, teachers and children. *Journal of Abnormal Child Psychology*, *38*(8), 1179–1191. https://doi.org/10.1007/S10802-010-9434-X

Goodman, R. (1997). The Strengths and Difficulties Questionnaire: A research note. *Journal of Child Psychology and Psychiatry*, *38*(5), 581–586. https://doi.org/10.1111/J.1469-7610.1997.TB01545.X

Goodman, R. (2001). Psychometric properties of the Strengths and Difficulties Questionnaire. *Journal of the American Academy of Child & Adolescent Psychiatry*, *40*(11), 1337–1345. https://doi.org/10.1097/00004583-200111000-00015

Guzman Holst, C., Bowes, L., Waite, P., Skripkauskaite, S., Shum, A., Pearcey, S., Raw, J., Patalay, P., & Creswell, C. (2023). Examining Children and adolescent mental health trajectories during the COVID-19 pandemic: Findings from a year of the Co-SPACE study. *JCPP Advances*, *3*(2), e12153. https://doi.org/10.1002/jcv2.12153

Hair, J. F., Ringle, C. M., & Sarstedt, M. (2011). PLS-SEM: Indeed a silver bullet. *Journal of Marketing Theory and Practice*, *19*(2), 139–152. https://doi.org/10.2753/MTP1069-6679190202

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*(3), 380–406. https://doi.org/10.1037/1045-3830.22.3.380

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8), 1–28. https://doi.org/10.18637/jss.v042.i08

Hu, Y., & Qian, Y. (2021). COVID-19 and adolescent mental health in the United Kingdom. *Journal of Adolescent Health*, *69*(1), 26–32. https://doi.org/10.1016/J.JADOHEALTH.2021.04.005

Ioannidis, J. P. A. (2014). How to make more published research true. *PLOS Medicine*, *11*(10), e1001747. https://doi.org/10.1371/JOURNAL.PMED.1001747

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling* (R package version 0.5-6).

Kai Yee, H., Bee Seok, C., & Ling, C. (2022). The factor structure among primary school children of the Strengths and Difficulties Questionnaire of Parents (SDQ-PR) in Malaysia during COVID-19. *Cogent Social Sciences*, *8*(1), 2126091. https://doi.org/10.1080/23311886.2022.2126091

Kankaanpää, R., Töttö, P., Punamäki, R. L., & Peltonen, K. (2023). Is it time to revise the SDQ? The psychometric evaluation of the Strengths and Difficulties Questionnaire. *Psychological Assessment*, *35*(12), 1069–1084. https://doi.org/10.1037/PAS0001265

Kerns, C. E., Elkins, R. M., Carpenter, A. L., Chou, T., Green, J. G., & Comer, J. S. (2014). Caregiver distress, shared traumatic exposure, and child adjustment among area youth following the 2013 Boston Marathon bombing. *Journal of Affective Disorders*, *167*, 50–55. https://doi.org/10.1016/J.JAD.2014.05.040

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

Kessler, R. C., Andrews, G., Colpe, L. J., Hiripi, E., Mroczek, D. K., Normand, S. L. T., Walters, E. E., & Zaslavsky, A. M. (2002). Short screening scales to monitor population prevalences and trends in non-specific psychological distress. *Psychological Medicine*, *32*(6), 959–976. https://doi.org/10.1017/S0033291702006074

Kline, R. B. (2015). *Principles and practice of structural equation modeling* (4th ed.). The Guilford Press. https://psycnet.apa.org/record/2015-56948-000

Knowles, G., Gayer-Anderson, C., Turner, A., Dorn, L., Lam, J., Davis, S., Blakey, R., Lowis, K., Pinfold, V., Creary, N., Dyer, J., Hatch, S. L., Ploubidis, G., Bhui, K., Harding, S., & Morgan, C. (2022). Covid-19, social restrictions, and mental distress among young people: A UK longitudinal, population-based study. *Journal of Child Psychology and Psychiatry*, *63*(11), 1392–1404. https://doi.org/10.1111/JCPP.13586

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/J.JCM.2016.02.012

Kostyrka-Allchorne, K., Creswell, C., Byford, S., Day, C., Goldsmith, K., Koch, M., Gutierrez, W. M., Palmer, M., Raw, J., Robertson, O., Shearer, J., Shum, A., Slovak, P., Waite, P., & Sonuga-Barke, E. J. S. (2021). Supporting Parents & Kids Through Lockdown Experiences (SPARKLE): A digital parenting support app implemented in an ongoing general population cohort study during the COVID-19 pandemic: A structured summary of a study protocol for a randomised controlled trial. *Trials*, *22*(1), 1–3. https://doi.org/10.1186/S13063-021-05226-4

Krueger, R. F., & Markon, K. E. (2006). Reinterpreting comorbidity: A model-based approach to understanding and classifying psychopathology. *Annual Review of Clinical Psychology*, *2*, 111–133. https://doi.org/10.1146/ANNUREV.CLINPSY.2.022305.095213

Lance, C. E., Butts, M. M., & Michels, L. C. (2006). The sources of four commonly reported cutoff criteria. *Organizational Research Methods*, *9*(2), 202–220. https://doi.org/10.1177/1094428105284919

Little, T. D. (2013). *Longitudinal structural equation modeling*. The Guilford Press.

Maassen, E., D'Urso, E. D., van Assen, M. A. L., Nuijten, M. B. M., De Roover, K., & Wicherts, J. M. (2023). The dire disregard of measurement invariance testing in psychological science. *Psychological Methods*, *30*(5), 966–979. https://doi.org/10.1037/MET0000624

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis-testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*(3), 320–341. https://doi.org/10.1207/S15328007SEM1103_2

Matsuishi, T., Nagano, M., Araki, Y., Tanaka, Y., Iwasaki, M., Yamashita, Y., Nagamitsu, S., Iizuka, C., Ohya, T., Shibuya, K., Hara, M., Matsuda, K., Tsuda, A., & Kakuma, T. (2008). Scale properties of the Japanese version of the Strengths and Difficulties Questionnaire (SDQ): A study of infant and school children in community samples. *Brain and Development*, *30*(6), 410–415. https://doi.org/10.1016/J.BRAINDEV.2007.12.003

Matsumoto, D., & Van de Vijver, F. J. R. (2010). *Cross-cultural research methods in psychology* (pp. 1–392). Cambridge University Press. https://doi.org/10.1017/CBO9780511779381

McMahon, J., March, S., Oakes, M., Silverman, W. K., Creswell, C., Rowe, A., Rajabi, M., & Skripkauskaite, S. (2025). Addressing international research challenges in child and adolescent mental health during global crises: Experience and recommendations of the Co-SPACE international consortium. *Child and Adolescent Psychiatry and Mental Health*, *19*(1), 62. https://doi.org/10.1186/S13034-025-00918-0

Mewton, L., Kessler, R. C., Slade, T., Hobbs, M. J., Brownhill, L., Birrell, L., Tonks, Z., Teesson, M., Newton, N., Chapman, C., Allsop, S., Hides, L., McBride, N., & Andrews, G. (2016). The psychometric properties of the Kessler Psychological Distress Scale (K6) in a general population sample of adolescents. *Psychological Assessment*, *28*(10), 1232–1242. https://doi.org/10.1037/pas0000239

Mieloo, C. L., Bevaart, F., Donker, M. C. H., Van Oort, F. V., Raat, H., & Jansen, W. (2013). Validation of the SDQ in a multi-ethnic population of young children. *European Journal of Public Health*, *24*(1), 26–32. https://doi.org/10.1093/EURPUB/CKT100

Moehring, A., Guertler, D., Krause, K., Bischof, G., Rumpf, H. J., Batra, A., Wurm, S., John, U., & Meyer, C. (2021). Longitudinal measurement invariance of the patient health questionnaire in a German sample. *BMC Psychiatry*, *21*(1), 1–10. https://doi.org/10.1186/S12888-021-03390-0/TABLES/7

Murray, A. L., Speyer, L. G., Hall, H. A., Valdebenito, S., & Hughes, C. (2021). 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*, *29*(6), 1248–1261. https://doi.org/10.1177/10731911211009312

Nunnally, J. C. (1978). An overview of psychological measurement. *Clinical Diagnosis of Mental Disorders*, 97–146. https://doi.org/10.1007/978-1-4684-2490-4_4

Oakes, M., Waite, P., Creswell, C., & Skripkauskaite, S. (2023). *Changes in children's mental health and parents' financial stress from March 2020 to March 2023* (Report 14).

Olino, T. M., Case, J. A. C., Hawes, M. T., Szenczy, A., Nelson, B., & Klein, D. N. (2021). Testing invariance of measures of internalizing symptoms before and after a major life stressor: The impact of COVID-19 in an adolescent and young adult sample. *Assessment*, *29*(7), 1371–1380. https://doi.org/10.1177/10731911211015315

Olino, T. M., Finsaas, M., Dougherty, L. R., & Klein, D. N. (2018). Is parent–child disagreement on child anxiety explained by differences in measurement properties? An examination of measurement invariance across informants and time. *Frontiers in Psychology*, *9*, 377776. https://doi.org/10.3389/FPSYG.2018.01295

Oort, F. J. (2005). Using structural equation modeling to detect response shifts and true change. *Quality of Life Research*, *14*(3), 587–598. https://doi.org/10.1007/S11136-004-0830-Y

Ortuño-Sierra, J., Fonseca-Pedrero, E., Aritio-Solana, R., Velasco, A. M., de Luis, E. C., Schumann, G., Cattrell, A., Flor, H., Nees, F., Banaschewski, T., Bokde, A., Whelan, R., Buechel, C., Bromberg, U., Conrod, P., Frouin, V., Papadopoulos, D., Gallinat, J., Garavan, H., . . .Lawrence, C. (2015). New evidence of factor structure and measurement invariance of the SDQ across five European nations. *European Child and Adolescent Psychiatry*, *24*(12), 1523–1534. https://doi.org/10.1007/S00787-015-0729-X

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review*, *41*, 71–90. https://doi.org/10.1016/J.DR.2016.06.004

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Revelle, W. (2023). *psych: Procedures for psychological, psychometric, and personality research* (2.3.6). Northwestern University. https://rdrr.io/cran/psych/

Ribeiro Santiago, P. H., Manzini, D., Haag, D., Roberts, R., Smithers, L. G., & Jamieson, L. (2022). Exploratory graph analysis of the Strengths and Difficulties Questionnaire in the longitudinal study of Australian children. *Assessment*, *29*(8), 1622–1640. https://doi.org/10.1177/10731911211024338

Robitzsch, A., & Lüdtke, O. (2023). Why full, partial, or approximate measurement invariance are not a prerequisite for meaningful and valid group comparisons. *Structural Equation Modeling: A Multidisciplinary Journal*, *30*(6), 859–870. https://doi.org/10.1080/10705511.2023.2191292

Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Runge, R. A., & Soellner, R. (2019). Measuring children's emotional and behavioural problems: Are SDQ parent reports from native and immigrant parents comparable? *Child and Adolescent Psychiatry and Mental Health*, *13*(1), 1–15. https://doi.org/10.1186/S13034-019-0306-Z

Rutkowski, L., & Svetina, D. (2013). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. https://doi.org/10.1177/0013164413498257

Sapin, A., Vanier, A., Descarpentry, A., Maffre Maviel, G., Vuillermoz, C., Falissard, B., Galera, C., Warszawski, J., Davisse-Paturet, C., Hazo, J. B., & Rouquette, A. (2024). Parental mental health and reporting of their child's behaviour: Measurement invariance of the French version of the parental Strengths and Difficulties Questionnaire. *European Child & Adolescent Psychiatry*, *33*, 3263–3272. https://doi.org/10.1007/S00787-024-02392-Z

Shibata, Y., Okada, K., Fukumoto, R., & Nomura, K. (2015). Psychometric properties of the parent and teacher forms of the Japanese version of the Strengths and Difficulties Questionnaire. *Brain and Development*, *37*(5), 501–507. https://doi.org/10.1016/J.BRAINDEV.2014.08.001

Smits, I. A. M., Theunissen, M. H. C., Reijneveld, S. A., Nauta, M. H., & Timmerman, M. E. (2016). Measurement invariance of the parent version of the Strengths and Difficulties Questionnaire (SDQ) across community and clinical populations. *European Journal of Psychological Assessment*, *34*(4), 238–246. https://doi.org/10.1027/1015-5759/A000339

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

Sourander, A., Westerlund, M., Kaneko, H., Heinonen, E., Klomek, A. B., How Ong, S., Fossum, S., Kolaitis, G., Lesinskiene, S., Li, L., Nguyen, M. H., Kumar Praharaj, S., Wiguna, T., Zamani, Z., & Gilbert, S. (2024). Cross-cultural comparison of the Strengths and Difficulties Self-Report Questionnaire in 12 Asian and European countries. *Journal of the American Academy of Child & Adolescent Psychiatry*, *64*(7), 799–809. https://doi.org/10.1016/J.JAAC.2024.10.002

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

Stevanovic, D., Urbán, R., Atilola, O., Vostanis, P., Singh Balhara, Y. P., Avicenna, M., Kandemir, H., Knez, R., Franic, T., & Petrov, P. (2015). Does the Strengths and Difficulties Questionnaire—Self report yield invariant measurements across different nations? Data from the International Child Mental Health Study Group. *Epidemiology and Psychiatric Sciences*, *24*(4), 323–334. https://doi.org/10.1017/S2045796014000201

Stone, L. L., Janssens, J. M. A. M., Vermulst, A. A., Van Der Maten, M., Engels, R. C. M. E., & Otten, R. (2015). The Strengths and Difficulties Questionnaire: Psychometric properties of the parent and teacher version in children aged 4–7. *BMC Psychology*, *3*(1), 4. https://doi.org/10.1186/s40359-015-0061-8

Stone, L. L., Otten, R., Engels, R. C. M. E., Vermulst, A. A., & Janssens, J. M. A. M. (2010). Psychometric properties of the parent and teacher versions of the Strengths and Difficulties Questionnaire for 4- to 12-year-olds: A review. *Clinical Child and Family Psychology Review*, *13*(3), 254–274. https://doi.org/10.1007/S10567-010-0071-2

Tashiro, A., & Shaw, R. (2020). COVID-19 pandemic response in Japan: What is behind the initial flattening of the curve? *Sustainability*, *12*(13), 5250. https://doi.org/10.3390/SU12135250

Toseeb, U., Oginni, O., Rowe, R., & Patalay, P. (2022). Measurement invariance of the Strengths and Difficulties Questionnaire across socioeconomic status and ethnicity from ages 3 to 17 years: A population cohort study. *PLOS ONE*, *17*(12), e0278385. https://doi.org/10.1371/JOURNAL.PONE.0278385

Turcan, C., Delamain, H., Loke, A., Pender, R., Mandy, W., & Saunders, R. (2024). Measurement invariance of the parent-reported Strengths and Difficulties Questionnaire in autistic adolescents. *Autism*, *28*(10), 2623–2636. https://doi.org/10.1177/13623613241236805

Usami, M., Iwadare, Y., Watanabe, K., Kodaira, M., Ushijima, H., Tanaka, T., Harada, M., Tanaka, H., Sasaki, Y., Okamoto, S., Sekine, K., & Saito, K. (2014). Prosocial behaviors during school activities among child survivors after the 2011 Earthquake and Tsunami in Japan: A retrospective observational study. *PLOS ONE*, *9*(11), e113709. https://doi.org/10.1371/JOURNAL.PONE.0113709

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement Invariance. *Frontiers in Psychology*, *6*, 1064. https://doi.org/10.3389/fpsyg.2015.01064

Vizard, T., Sadler, K., Ford, T., Newlove-Delgado, T., McManus, S., Marcheselli, F., Davis, J., Williams, T., Leach, C., Mandalia, D., & Cartwright, C. (2020). *Mental health of children and young people in England, 2020* (Issue July). https://files.digital.nhs.uk/AF/AECD6B/mhcyp_2020_rep_v2.pdf

Waite, P., Pearcey, S., Shum, A., Raw, J. A. L., Patalay, P., & Creswell, C. (2021). How did the mental health symptoms of

children and adolescents change over early lockdown during the COVID-19 pandemic in the UK? *JCPP Advances*, *1*(1), 1–10. https://doi.org/10.1111/jcv2.12009

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. https://doi.org/10.1111/J.1750-8606.2009.00110.X

Woerner, W., Fleitlich-Bilyk, B., Martinussen, R., Fletcher, J., Cucchiaro, G., Dalgalarrondo, P., Lui, M., & Tannock, R. (2004). The Strengths and Difficulties Questionnaire overseas: Evaluations and applications of the SDQ beyond Europe. *European Child and Adolescent Psychiatry, Supplement*, *13*(2), ii47–ii54. https://doi.org/10.1007/S00787-004-2008-0

Zhang, X., Lv, T., Leavey, G., Zhu, N., Li, X., Li, Y., & Chen, Y. (2023). Does depression affect the association between prosocial behavior and anxiety? A cross-sectional study of students in China. *Frontiers in Public Health*, *11*, 1274253. https://doi.org/10.3389/FPUBH.2023.1274253