ORIGINAL ARTICLE

Review of the strengths and difficulties questionnaire translated into languages spoken by children and adolescents of refugee background

Yvonne Stolk | Ida Kaplan | Josef Szwarc

Victorian Foundation for Survivors of Torture,
Brunswick, Victoria, Australia

Correspondence

Yvonne Stolk, Research Consultant, Victorian
Foundation for Survivors of Torture, 4
Gardiner Street, Brunswick, VIC 3056,
Australia
Email: yvonne.stolk@bigpond.com.au

Abstract

The Strengths and Difficulties Questionnaire (SDQ), routinely used to screen for children's and adolescents' emotional and behavioural problems, has been translated into at least 80 languages. As children of refugee background are particularly vulnerable to mental health problems resulting from their refugee experiences, this review examines whether SDQs translated into languages spoken by major refugee groups are validated and culturally equivalent to the original SDQ and sensitive to change following interventions. No reviewed studies of translated SDQs reported on challenges in achieving conceptual and linguistic equivalence in translation. Factor analysis predominantly showed structural inequivalence with the original 5-factor model, suggesting translated SDQ subscales may measure different constructs. Predictive equivalence findings tended to show somewhat higher sensitivity for detecting disorder than the original SDQ's low sensitivity, and somewhat lower specificity. Outcome studies yielded equivocal results with refugee and immigrant groups. SDQ items do not detect the psychological sequelae of trauma; hence it is recommended that the SDQ be used with caution to screen refugee children, with a follow-up clinical interview for verification. Cross-cultural qualitative research is needed into parents' and adolescents' interpretation of translated SDQ items.

KEYWORDS

child and adolescent, cultural equivalence, refugee, Strengths and Difficulties Questionnaire (SDQ), translation

1 | INTRODUCTION

The Strengths and Difficulties Questionnaire (SDQ) is widely used internationally to identify children's and adolescents' emotional, behavioural, and relationship problems and prosocial behaviour (Du, Kou, & Coghill, 2008; Goodman, 1997; Goodman, Meltzer, & Bailey, 1998; Stevanovic et al., 2014; Vostanis, 2006). Its brief one-page format and inclusion of positive attributes has facilitated the SDQ's adoption as a screening instrument in national mental health surveys in the United Kingdom (Goodman, 2001), the United States (Bourdon, Goodman, Rae, Simpson, & Koretz, 2005), and Australia, replacing the longer Child Behaviour Checklist (CBCL; Lawrence et al., 2015). The SDQ is variously recommended as a useful part of clinical assessment of child psychopathology, psychological adjustment (Goodman, 2001), psychiatric disorders (Goodman, Ford, Simmons, Gatward, & Meltzer, 2003), and mental health (Goodman, Lamping, & Ploubidis,

2010). The United Kingdom (UK) and Australian mental health services use the SDQ as an outcome measure (Australian Mental Health Outcomes and Classification Network, 2015; Fugard et al., 2015), as have specialist programmes in clinics (Lee, Jones, Goodman, & Heyman, 2005), and schools (Rousseau et al., 2007).

The mental health of children of refugee background (hereafter referred to as "refugee children") has also been investigated using the SDQ (Derluyn & Broekaert, 2007; Lorek et al., 2009), but whether SDQs translated from the original English version are validated and culturally appropriate screening instruments for non-English-speaking refugee children and parents, has not been investigated. Children and adolescents aged under 18 years comprised 51% of the world's 21.3 million refugees in 2015 (United Nations High Commissioner for Refugees (UNHCR), 2015). Refugee children are particularly vulnerable to mental health problems resulting from multiple stressors associated with the refugee experience, including forced displacement, complex

trauma, separation from parents and multiple other losses, time in refugee camps, detention on arrival, and resettlement adjustments in the host country (Henley & Robinson, 2011; Kaplan, 2013). A review of refugee children's mental health in Western settlement countries found predominantly elevated prevalence of disorders, with anxiety ranging 33–50%, depression 3–30% and post-traumatic stress disorder (PTSD) 19–54%; which compares to a general child population PTSD prevalence of 2 to 9% (Bronstein & Montgomery, 2011). Worldwide data on the psychological problems of non-traumatized children and adolescents are scarce, but prevalence has been estimated at 10–20% (Kielsing et al., 2011).

These prevalence findings bring into sharp focus the need for mental health screening of the thousands of refugee children and adolescents fleeing the current conflicts in the Middle East and other countries, and being resettled in Western countries (Hebebrand et al., 2016). It is imperative however, to determine whether instruments presently being used (including the SDQ) have evidence for validity with young refugees (Gadeberg & Norredam, 2016).

Timely use of validated screening instruments may aid in identifying and preventing development of more serious disorders (Gadeberg & Norredam, 2016; Stone, Otten, Engels, Vermulst, & Janssens, 2010). However, refugee respondents are likely to require translated SDQs as the majority of refugee families originate from countries where English is not the first language (Refugee Council of Australia (RCA), 2016a), while more than 80% of 73,000 refugees resettled in 2014, were permanently resettled in English-speaking countries (UNHCR, 2016). As of July 2015 at least 80 SDQ translations were available cost-free on the SDQ website (youthinmind, 2015), not all of which are validated.

Several authors have examined evidence for the SDQ's validity and reliability across various countries (Achenbach et al., 2008; Goodman et al., 2012; Stevanovic et al., 2014; Stone et al., 2010), but these authors did not report languages spoken by participants, nor whether conceptual equivalence was achieved in translation. Stone et al. (2010) combined studies conducted in English- and non-English-speaking countries to reach overall conclusions about the SDQ's validity, thereby obscuring potential sources of psychometric variation due to the effects of translation (Van de Vijver & Leung, 2011).

This article aims to review the cultural equivalence of SDQs translated into languages spoken by major refugee groups resettled in Western countries in the past 10 years. To establish cultural equivalence, evidence for the original SDQ's validity and reliability is first examined enabling comparisons with translated SDQs. As refugee children may have experienced trauma, the SDQ's responsiveness to the psychological sequelae of exposure to traumatic events is examined, and as refugee children are the subject of outcome studies, the SDQ's sensitivity to change with refugee children is reviewed.

2 | METHOD

2.1 | Language selection

To determine which SDQ translations were in languages spoken by major refugee groups arriving in Western settlement countries, we

drew on UNHCR refugee resettlement data between 2004 and 2015, compiled by the RCA (2016b), and the UK Refugee Council's (2015) language profiles for major refugee groups. Of the 80 SDQ translations, 16 were in languages spoken in the main source countries for resettled refugees in the past 10 years. The languages are listed here in alphabetical order with the refugee-source countries in parentheses: Amharic (Ethiopia), Arabic (Iraq, Sudan, Syria, Yemen), Chinese (China), Chin Haka (Myanmar/Burma), Dari and Pashtu (Afghanistan), Farsi/Persian (Iran), French [Democratic Republic of the Congo (DR of Congo)], Khmer (Cambodia), Kurdish (Iraq, Turkey), Russian (Russian Federation), Serbian (Serbia), Somali (Somalia), Tamil (Sri Lanka), Turkish (Turkey), and Urdu (Pakistan).

2.2 | Study selection criteria

DISCOVERY EBSCOhost, Medline, PsycINFO, CINAHL (Cumulative Index of Nursing and Allied Health Literature) and Academic Search Complete databases were searched from 1994 to July 2015 for peer-reviewed, English-language publications. Searches for reports on the original SDQ's development and validation used the terms, "Strengths and Difficulties Questionnaire" (2,018 studies), AND valid* (199). Following review of abstracts 169 studies were eliminated, including duplicate publications; validations of non-refugee languages, studies using the SDQ to validate another measure; and conference abstracts. Searches relating to validation of SDQs translated into languages spoken in refugee-source countries used combinations of the terms: refugee*, asylum, ethnic, translat*; cultural equiv*, valid*; outcome, sensitivity to change, intervention, and each of the refugee languages listed earlier. Additional searches were conducted on the internet, youthinmind (2012a) and in reference lists of selected papers. A total of 34 studies were retained.

2.3 | Validity and cultural equivalence criteria

Translation effectively creates a new instrument that demonstrates cultural equivalence by providing evidence of validity and reliability comparable to the original instrument (Van de Vijver & Leung, 2011). Construct validity refers to how well a test measures the construct it was designed to measure, and how well theory and evidence support test score interpretations. Evidence for construct validity may come from one or more of: test content, response processes (obtained through participant cognitive interview), internal structure, and relations to other variables, supported by standardized administration and norm-based score interpretation guidelines (American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education, 2014).

To establish cultural equivalence translated instruments test for constructs that have similar meaning across cultural groups, providing evidence of at least one type of equivalence: conceptual and linguistic; method and administrative; structural; item and scalar; criterion or predictive; and normative equivalence (Okawa, 2008; Van de Vijver & Leung, 2011).

3 | RESULTS

The 34 selected studies comprised seven validations of the original SDQ, 17 validations of SDQs translated into 11 languages spoken in refugee source-countries (hereafter called refugee-relevant languages), three validations with immigrant groups from refugee-source countries, and seven sensitivity-to-change studies with refugee and immigrant groups. It should be noted that no studies were found that validated SDQs with children and adolescents identified as refugees. Consequently we selected studies that validated the SDQ in refugee-source-countries (e.g. Yemen), as these translations are available online (youthinmind, 2015). The three validations with immigrant groups included children from refugee source-countries. We use the terms ethnic and immigrant groups inter-changeably, unless refugees are specified.

The validation studies were of SDQs translated into: Arabic (three studies), Chinese (four studies), Dari and Pashtu (one study), Farsi/Persian (two studies), French (one study), Russian (three studies), Tamil (one study), Turkish and Serbian (one study, included in a cross-national validation), and Urdu (one study). No peer reviewed English-language validation studies were found for five of the 16 refugee-relevant translations on the SDQ website: Amharic, Chin Haka, Khmer, Kurdish, and Somali. Table 1 summarizes reviewed studies with participant details, and the psychometric properties investigated. Subsequent tables (Tables 2–7) report results of those investigations.

3.1 | The original SDQ

The 25-item SDQ was developed to provide a briefer scale than the Rutter Parent Questionnaire (RPQ; Elander & Rutter, 1996) and CBCL (Achenbach, 1991), and to include strengths as well as weaknesses (Goodman, 1997). The similarly-worded parent, teacher and self-rated SDQs (P-, T-, and S-SDQ, respectively) comprise four 5-item problem subscales (emotional symptoms, conduct problems, hyperactivity/inattention, peer problems), and a positively-worded prosocial behaviour subscale, designed to improve acceptability (Goodman, 1997; Goodman et al., 1998). Five positively-worded reverse-scored items are also included in the problem subscales. Response scales are: 0, Not true; 1, Somewhat true; 2, Certainly true. Only the four problem subscales scores are summed, yielding a maximum total difficulties score (TDS) of 40 (Goodman, 1997). The P- and T-SDQs were recently described as suitable for 4–17 year-olds, and the S-SDQ for 11–17 year-olds (youthinmind, 2015).

Cutoff scores (see Table 8) designate scores below the 80th percentile in the “normal” range, the 80th to 90th percentile in the “borderline”, and above the 90th percentile in the “abnormal” range (Goodman, 1997; Goodman et al., 1998). Goodman (1997) indicated that cutoffs would be expected to vary with a population's level of risk, gender and age; ethnicity was not mentioned.

An SDQ impact supplement asks whether the child has difficulties, and about associated chronicity, distress, social impairment, and burden to others (Goodman, 1999). The impact supplement's psychometric properties are not further discussed as it was examined by only one translation study.

3.1.1 | Validity and reliability of the SDQ

SDQ item content and subscales were based on factor analysis of the well-established RPQ, with which it showed high correlations. Items selected aimed to represent key DSM-IV (American Psychiatric Association, 1994) symptoms (Goodman & Scott, 1999). Scales were modified following “advice from colleagues” and “informal trials” (Goodman, 1997, p. 582). Goodman (2001) conducted factor analysis (method not reported), choosing a 5-factor solution (Table 2). All items loaded on predicted factors, although up to eight items cross-loaded ≥ 0.30 . Four T- and S-SDQ positively-worded items cross-loaded on the prosocial factor, which Goodman (2001) also interpreted as a positive construal factor.

Internal consistency across the three informants was highest for the TDS and hyperactivity subscale, and poorest for peer problems (Table 3). Test-retest stability after 4–6 months, acknowledged as too long for stability, showed low to acceptable correlations (Table 3; Goodman, 2001). Internal convergent validity was evident in high intercorrelations between the conduct and hyperactivity subscales, and the emotional and peer problems subscales. Lower correlations between the emotional and externalizing subscales provided evidence for discriminant validity (Table 4; Goodman, 2001). Confirmatory factor analysis (CFA) by Goodman et al. (2010) found the high subscale intercorrelations consistent with a 3-factor second-order model (externalizing, internalizing and prosocial behaviour) providing a better fit than the first-order model (Table 2). Nevertheless, the five subscales are recommended for screening clinical samples, and the second-order subscales for epidemiological studies (Goodman et al., 2010; youthinmind, 2012b).

Informant intercorrelations between parent/teacher, parent/self and teacher/self SDQ TDSs (Goodman, 2001; Table 5) were higher than respective meta-analytic means of $r = 0.27, 0.25$ and 0.20 , established by Achenbach, McConaughy, and Howell's (1987) meta-analysis of cross-informant correlations. Across informant pairs intercorrelations were strongest on the hyperactivity subscale and TDS, and weakest on the prosocial and emotional subscales.

Evidence for the P-SDQ's external convergent validity came from moderate to high correlations with comparable subscales of the RPQ (Goodman, 1997), and CBCL (Table 6; Goodman & Scott, 1999).

Providing evidence for predictive validity, the S-SDQ differentiated small samples of clinic from community youth with satisfactory to low areas under the curves (AUCs) (Table 7; Goodman et al., 1998). In a national survey children with “high risk” SDQ scores (> 90 th percentile) were more likely than a “low risk” group to be diagnosed with any DSM-IV disorder on the Development and Well-being Assessment (DAWBA; Goodman, Ford, Richards, Gatward, & Meltzer, 2000; Goodman, 2001). Mean odds ratios ranged from P-SDQ 15.7 to S-SDQ 6.2 (Table 7; Goodman, 2001). Prediction of anxiety disorders from the emotional subscale was least accurate, with T- and S-SDQ sensitivity of 29%, while prediction of attention deficit hyperactivity disorder (ADHD) from the hyperactivity subscale was most accurate with P-SDQ sensitivity of 74%, and T-SDQ 68%. Specificity exceeded 90% across diagnoses and informants (Goodman, 2001).

A multi-informant algorithm (M-IA), which draws on symptom and impact scores from up to three informants, improved the SDQ's predictive accuracy for diagnoses. Sensitivities for any DAWBA-based

TABLE 1 Summary of reviewed studies of the original and translated Strengths and Difficulties Questionnaires (SDQs)

Author, Study purpose	Language, Country ^a	Sample source, informants, N ^a	Age (years)	Translation protocol	Validity evidence evaluated
<i>Reference studies</i>					
Goodman, 1997 Validate P and T-SDQ	English, UK	Psychiatric/Dental P 188/158; T 146/39	4–16	NA	Interrater, subscale correlations, convergent, predictive validity, means, cutoffs
Goodman et al., 1998 Validate S-SDQ	English, UK	CAMHS/Community P 54/84; T 54/54; S 116/83	11–16	NA	Internal consistency, interrater correlations, predictive validity, means, cutoffs
Goodman, 1999 Validate impact supplement	English, UK	CAMHS/Community P 467/232; T 331/121; S 186/89	5–15	NA	Test-retest, interrater, convergent, predictive validity, means, cutoffs
Goodman & Scott, 1999 SDQ/CBCL evidence for convergent, predictive validity	English, UK	Psychiatric/Dental P 61/71	4–7	NA	Convergent, predictive validity, means
Goodman, 2001 Validate P, T, S-SDQ	English, UK	1999 British National Survey/Retest P 9998/2091; T 7313/796; S 3983/781 P/T/S DAWBA diagnosis 500	P, T 5–15 S 11–15	NA	Factor analysis, ^b internal consistency, test-retest reliability, interrater correlations, convergent, predictive, means, cutoffs
Goodman et al., 2003 Validate SDQ multi-informant algorithm	English, UK	1999 British National Survey P/T/S 7,984	M = 10.2 (3.1)	NA	Predictive validity
Goodman et al., 2010 Investigate 2nd order-factors	English, UK	1999 & 2004 British National Surveys P 18,222; T 14,263; S 7,678	5–16	NA	CFA, internal consistency, subscale intercorrelations, predictive validity
<i>Translation studies</i>					
Thabet et al., 2000 Arabic P-SDQ factor structure	Arabic, Gaza strip	School children P 3 yrs N = 92; P/T 6 yrs N = 81 P/T 11 yrs N = 92; P/S 16 yrs N = 57. Total N = 322	3–16	FBT, ER, consultation with R. Goodman	CFA, internal consistency, interrater correlations, means
Almagrabi & Shuwail, 2004 Validate Arabic S-SDQ	Arabic, Yemen	Clinic/Schools S 57/600	12–17	FBT, ER, pilot-test	PCA, predictive validity, means, cutoffs
Alvahri & Goodman, 2006 Validate Arabic P, T-SDQ	Arabic, Yemen	Clinics/Schools P 87/100; T 68/100	5–12	SDQ revision, BT, consultation with R. Goodman	Predictive validity
Du et al., 2008 Validate Chinese P, T, S-SDQ	Chinese, Shanghai, China	ADHD patients/Controls/Schools P/T 47/48/1,965; P/T/S 47/48/690 P/T 6-week retest sub-sample: 45	3–17	FBT by Chinese & US academics	PCA, internal, test-retest reliability, subscale, interrater correlations, convergent, predictive validity, means, cutoffs
Lai et al., 2010 Validate re-translated Chinese P, T-SDQ	Chinese, Hong Kong	Clinic/Schools P 450/3,634; T 262/3,647 2–4 weeks school retest: P 471, T 732	6–12	FBT, ER	Internal consistency, test-retest reliability, subscale, interrater correlations, predictive validity, means

(Continues)

TABLE 1 (Continued)

Author, Study purpose	Language, Country ^a	Sample source, informants, N ^a	Age (years)	Translation protocol	Validity evidence evaluated
Liu et al., 2013 Validate Chinese P, T, S-SDQ	Chinese, Taiwan	ADHD clinic/Schools P190/3,546; T 200/3,669; S 76/2,672	6–15	NR	PCA, internal consistency, test–retest reliability, interrater correlations, convergent validity, means, cutoffs
Yao et al., 2009 Validate Chinese S-SDQ	Chinese, Hunan Province, China	Schools S 1,135 (Girls 574; Boys 561); 8-week retest 157	11–18	SDQ website	Factor analysis, internal consistency, test–retest reliability, subscale intercorrelations, convergent validity, means
Panther-Brick et al., 2011 Longitudinal mental health study	Dari, Pashtu, Afghanistan	Schools: Baseline/F-U P/S 364/234	11–16	FBT, focus groups, ER, pilot studies	Internal, test–retest reliability, interrater correlations, convergent validity, means
Kashala et al., 2005 Validate French T-SDQ	French, DR of Congo	Schools T 1,187	7–9	NR, French website version used	PCA, internal consistency, means, cutoffs
Ghanizadeh et al., 2007 Validate Persian P, T, S-SDQ	Persian/Farsi, Iran	Schools P/T 756; S 330; Clinical interview: 155	3–18	SDQ website	Factor analysis, internal consistency, interrater correlations, predictive validity
Shahrivar et al., 2009 Validate Persian P, T-SDQ	Persian/Farsi, Iran	Schools P/T 600; Clinical interview: 52	6–12	Translation under supervision of R. Goodman	Internal consistency, subscale intercorrelations, predictive validity
Goodman et al., 2005 SDQ M-IA prevalence of child mental health problems	Russian, Russia	Schools P 448; T/S N not specified Diagnosis: screen positive 87; screen negative 85	7–14	Reported in Russian-language publication	Predictive validity
Ruchkin et al., 2007 Validate Russian S-SDQ	Russian, Russia	Schools S 2,892	13–18	FBT, discrepancies corrected	CFA, internal consistency, subscale intercorrelations, means
Ruchkin et al., 2012 Validate Russian T-SDQ	Russian, Russia	Schools T 538; S 2,892	13–18	Reported in Russian-language publication	CFA, internal consistency, subscale, interrater correlations, convergent validity, means
Lukumar et al., 2008 Validate Tamil S-SDQ	Tamil, Sri Lanka	Clinic/Community controls S 35/91 School sample to estimate internal consistency: S 1,199	11–16	FBT, consensus resolution	Internal consistency, subscale intercorrelations, predictive validity
Samad et al., 2005 Validate Urdu P-SDQ	Urdu, Pakistan	Psychiatric/Paediatric clinic P 72/140	4–16	FBT. Oral administration to illiterate parents	Predictive validity, means, cutoff
Stevanovic et al., 2014 S-SDQ factor invariance across seven countries	Languages NR 7 countries →	International Mental Health Study Group: Bulgaria 265, Croatia, 293, India 393, Indonesia 228, Nigeria 522, Serbia 386 , Turkey 280 , Total 2367	13–18	NR. Directed to SDQ website	CFA, EFA, means
<i>Refugee and immigrant groups</i>					
Richter et al., 2011 Compare S-SDQ factor structure for Norwegian, refugee and immigrant youth	Norwegian, other languages NR. Pakistan , Iran , Turkey , Somalia , Vietnam. Norway	Oslo Health Study: School children: Norway 5,379 Pakistan 516 Other ethnic (Iran 89, Turkey 93, Somalia 76, Vietnam 91) total 349	15–16	Norwegian SDQ translated into English for non-Norwegian speakers; assisted by aide.	CFA, internal consistency, subscale intercorrelations

(Continues)

TABLE 1 (Continued)

Author, Study purpose	Language, Country ^a	Sample source, informants, N ^a	Age (years)	Translation protocol	Validity evidence evaluated
Zwirs et al., 2011 Validate T-SDQ in multi-ethnic population	Dutch, Morocco, Turkey, Surinam, Netherlands	Schools, T-SDQ: Dutch 684, Moroccan 702, Turkish 434, Surinamese 365, Total 2,185	6–10	Validated Dutch translation	CFA, internal consistency, subscale intercorrelations, means
Mieloo et al., 2014 Validate P, T-SDQ in multi-ethnic population	Dutch, Surinamese, Antillean/Aruban, Arabic/Moroccan, Turkish, Netherlands	Child health care program Dutch P 4,750, T 4,516; Moroccan P 623, T 811; Turkish P 661, T 759; Other P 2066, T 2439 Totals P 8114, T 9355; P & T 6,525	5–6	Dual language Dutch/Arabic, Dutch/Turkish SDQs for Moroccan and Turkish parents. Translation process NR	PCA, internal consistency, interrater correlations convergent validity, means, cutoffs

^aIn multi-country or multi-ethnic studies languages and countries in bold typeface are refugee languages or source countries.

^bFactor analytic method not reported.

ADHD = attention deficit hyperactivity disorder, BT = back-translation, CAMHS = child and adolescent mental health service, CFA = confirmatory factor analysis; DAWBA = Development and Well-being Assessment structured interview, DR of Congo = Democratic Republic of Congo, ER = expert review, EFA = exploratory factor analysis; FBT = independent forward- and back-translation, M-IA = multi-informant algorithm, NA = not applicable, NR = not reported, PCA = principal components analysis; P = parent, S = self, SDQ = Strengths and Difficulties Questionnaire, T = teacher.

diagnoses approximately doubled, compared to sensitivities for individual informants' SDQ TDS (Goodman et al., 2003).

3.2 | Translated SDQs and use with ethnic groups in settlement countries

This section reviews the 20 SDQ validation studies for 11 refugee-relevant translations (hereafter simply called translations) on the SDQ website (Arabic, Chinese, Dari and Pashtu, Farsi/Persian, French, Russian, Serbian, Tamil, Turkish and Urdu, including three validations with ethnic groups) to evaluate their cultural equivalence.

3.2.1 | Conceptual and linguistic equivalence

A conceptually and linguistically equivalent translated instrument has the same familiarity, meaning, and relevance, and makes identical performance demands in both languages [Arnold & Matus, 2000; International Test Commission (ITC), 2016]. Translations consequently need to adapt for cross-cultural differences in mental health concepts and idioms of distress; literacy levels; concordance of spoken and written language; and taboo topics (Acquadro, Conway, Hareendran, & Aaronson, 2008; Andary, Stolk, & Klimidis, 2003).

Translation protocols recommend a multi-step translation process. The instrument is forward- and back-translated by independent bilingual translators familiar with the field, in consultation with a bilingual expert advisory panel. Field-testing through cognitive interviews is undertaken with a respondent sub-sample to elicit their understanding of items, and ensure identical constructs are measured in both languages. Documentation is provided on translation and adaptation procedures, translators' background, and sampled test takers (Acquadro et al., 2008; ITC, 2016; World Health Organization (WHO), 2015).

Nine of the 20 validation studies reported variations on forward- and back-translation by independent expert translators, and expert panel review (Table 1). Two studies also reported wording modifications following pilot studies: the Arabic S-SDQ (Alyahri & Goodman, 2006), and Dari and Pashtu P- and S-SDQ (Panter-Brick, Goodman, Tol, & Eggerman, 2011). Translators of the Tamil (Lukumar, Wijewardana, Hermansson, & Lindmark, 2008) and Urdu (Samad, Hollis, Prince, & Goodman, 2005) SDQs endeavoured to maintain semantic and conceptual equivalence, with the latter subtly modifying potentially offensive words, such as *steals* or *lies*. In three studies translation occurred under the supervision of, or in consultation with R. Goodman, detail of which was not reported. No study reported investigating how well terminology translated, or whether translation challenges were encountered in attaining conceptual equivalence.

Three studies reported obtaining translations from the SDQ website without further explanation. Three did not report translation processes at all, including a cross-national SDQ comparison by Stevanovic et al. (2014), who stated that S-SDQ translations on the website are validated according to procedures specified by Goodman (2001). Such procedures are not specified in that paper and not all translated S-SDQs are validated.

3.2.2 | Method equivalence

Method and administrative equivalence require that administration is not influenced by cultural and linguistic factors, such as lack of

TABLE 2 Comparative factor loadings for the original and translated parent, teacher and self-rated Strengths and Difficulties Questionnaires (SDQs)

SDQ factors and items	P-SDQ				T-SDQ				S-SDQ			
	English, UK		Translations		English, UK		Translations		English, UK		Translations	
	Goodman, 2001	Goodman et al., 2010 ^a	Weighted M	Range of loadings ^c	Goodman, 2001	Goodman et al., 2010	Weighted M	Range of loadings	Goodman, 2001	Goodman et al., 2010	Weighted M	Range of loadings
	N = 9,998	N = 18,222	Total N = 7,117 ^b		N = 7,313	N = 14,263	Total N = 10,684 ^b		N = 3,983	N = 7,678	Total N = 9,060 ^b	
1. Emotional symptoms												
3. Somatic	0.47	0.46	0.37	0.00–0.81	0.53	0.64	0.44	0.32–0.56	0.41	0.48	0.49	0.37–0.61
8. Worries	0.69	0.68	0.60	0.46–0.69	0.77	0.78	0.67	0.57–0.77	0.74	0.66	0.67	0.45–0.75
13. Unhappy	0.60	0.86	0.49	0.07–0.53	0.64	0.92	0.44	0.00–0.96	0.51 ^{x2,x4}	0.77	0.70	0.64–0.80
16. Clingy	0.66	0.60	0.66	0.55–0.68	0.75	0.77	0.69	0.48–0.80	0.64	0.56	0.49	0.37–0.68
24. Fears	0.71	0.70	0.65	–0.09–0.73	0.80	0.84	0.73	0.61–0.78	0.72	0.67	0.59	0.46–0.68
2. Conduct problems												
5. Temper	0.54	0.67	0.46	0.00–0.74	0.63	0.77	0.60	0.00–0.77	0.52	0.66	0.37	0.00–0.66
7. Obedient	–0.43 ^{x3,x5}	–0.71	0.13	0.00–0.43	–0.37 ^{x3,x5}	–0.82	0.27	0.00–0.82	–0.36 ^{x5}	–0.59	0.07	–0.39–0.42
12. Fights, bullies	0.61	0.73	0.48	0.00–0.64	0.68	0.87	0.71	0.47–0.86	0.58	0.59	0.39	0.00–0.65
18. Lies, cheats	0.64	0.72	0.30	0.00–0.58	0.71	0.86	0.61	0.35–0.85	0.63	0.70	0.39	0.00–0.62
22. Steals/spiteful	0.52	0.68	0.31	–0.05–0.62	0.60	0.71	0.48	0.00–0.77	0.57	0.59	0.34	0.00–0.61
3. Hyperactivity												
2. Restless	0.66	0.73	0.48	0.00–0.78	0.79	0.90	0.65	0.00–0.93	0.77	0.56	0.52	0.00–0.69
10. Fidgety	0.65	0.78	0.22	0.00–0.66	0.80	0.91	0.63	0.00–0.89	0.73	0.65	0.35	0.00–0.73
15. Distractible	0.77	0.80	0.62	0.00–0.72	0.81	0.90	0.75	0.63–0.87	0.52 ^{x2}	0.74	0.57	0.00–0.71
21. Reflective	–0.64	–0.69	0.54	–0.52–0.68	–0.61 ^{x5}	–0.88	0.64	–0.52–0.86	–0.42 ^{x5}	–0.59	0.18	–0.41–0.60
25. Good attention	–0.72	–0.75	0.58	–0.41–0.73	–0.72 ^{x5}	–0.88	0.67	–0.55–0.82	–0.45 ^{x5}	–0.65	0.21	–0.26–0.65
4. Peer problems												
6. Rather solitary	0.56	0.50	0.01	0.00–0.12	0.69	0.54	0.20	0.00–0.63	0.54	0.47	0.37	0.00–0.75
11. Has good friend	–0.64	–0.67	0.01	0.00–0.28	–0.65	–0.80	0.23	–0.62–0.75	–0.46	–0.34	0.04	–0.57–0.51
14. Popular	–0.61	–0.82	0.02	0.00–0.52	–0.48 ^{x5}	–0.97	0.15	–0.83–0.98	–0.33 ^{x5}	–0.58	0.04	–0.41–0.49
19. Bullied	0.47 ^{x1}	0.67	0.03	0.00–0.63	0.45 ^{x1}	0.58	0.28	0.00–0.75	0.48 ^{x1}	0.73	0.29	0.00–0.84
23. Best with adults	0.57	0.49	0.08	–0.28–0.00	0.74	0.40	0.28	0.00–0.71	0.62	0.30	0.13	0.00–0.35
5. Prosocial												
1. Considerate	0.58 ^{x2}	0.82	0.53	0.21–0.58	0.68 ^{x2,x3}	0.92	0.35	–0.68–0.96	0.55 ^{x2}	0.76	0.56	0.37–0.70
4. Shares	0.53	0.71	0.61	0.54–0.65	0.63	0.80	0.44	–0.72–0.84	0.60	0.56	0.46	0.28–0.59
9. Caring	0.67	0.66	0.68	0.57–0.71	0.76	0.85	0.52	–0.78–0.84	0.65	0.66	0.49	0.00–0.78

(Continues)

TABLE 2 (Continued)

SDQ factors and items	P-SDQ				T-SDQ				S-SDQ			
	English, UK		Translations		English, UK		Translations		English, UK		Translations	
	Goodman, 2001	Goodman et al., 2010 ^a	Weighted M	Range of loadings ^c	Goodman, 2001	Goodman et al., 2010	Weighted M	Range of loadings	Goodman, 2001	Goodman et al., 2010	Weighted M	Range of loadings
	N = 9,998	N = 18,222		Total N = 7,117 ^b	N = 7,313	N = 14,263		Total N = 10,684 ^b	N = 3,983	N = 7,678		Total N = 9,060 ^b
17. Kind to kids	0.56	0.68	0.63	0.58–0.64	0.75	0.80	0.46	–0.70–0.77	0.60	0.66	0.53	0.41–0.64
20. Helps out	0.68	0.52	0.63	0.49–0.66	0.76	0.69	0.46	–0.71–0.75	0.65	0.59	0.60	0.49–0.70

Note: Included translation studies: Arabic P-SDQ, Thabet et al., 2000; Chinese P-, T-, S-SDQ Du et al., 2008; Chinese P-, T-, S-SDQ Liu et al., 2013; French, DR Congo T-SDQ Kashala et al., 2005; Netherlands Moroccan/Arabic, Turkish P-, T-SDQ, Mieloo et al., 2014; Netherlands Moroccan/Arabic, Turkish T-SDQ Zwirs et al., 2011; Norway, Pakistani minority, Other minority S-SDQ, Richter et al., 2011; Russian S-, T-SDQ Ruchkin et al., 2007, 2012; Serbia, Turkey S-SDQ Stevanovic et al., 2014. k = 10 (three studies did not report loadings).

^aFor Goodman et al. (2010) italic typeface items comprise second-order internalizing factor, bold typeface items second-order externalizing factor, other items prosocial factor.

^bSee Table 1 for N in individual studies.

^cZero loadings indicate item did not load on that factor in at least one study.

^{x1}Denotes item cross-load ≥ 0.30 on factor 1.

^{x2}Denotes item cross-load ≥ 0.30 on factor 2.

^{x3}Denotes item cross-load ≥ 0.30 on factor 3.

^{x4}Denotes item cross-load ≥ 0.30 on factor 4.

^{x5}Denotes item cross-load ≥ 0.30 on factor 5.

familiarity with test-taking or poorly translated instructions, but elicits the intended response across languages (ITC, 2016; Van de Vijver & Leung, 2011). To minimize method inequivalence manuals for translated instruments should provide cross-culturally consistent administration, scoring, and score interpretation guidelines, and information on administering tests with interpreters (Van de Vijver & Poortinga, 2005). The SDQ website provides English-language advice on scoring and score interpretation, but not on test administration; nor does this information appear to be available in non-English languages. Non-English-speaking test administrators may be unaware of appropriate administration procedures, potentially biasing results.

The majority of translation studies reported administration procedures: the P- and S-SDQ were sent home, or administered at school, but no studies reported on administration of the T-SDQ. While the SDQ is usually self-completed, in three studies interviewers administered P-SDQs to illiterate parents (Alyahri & Goodman, 2006; Samad et al., 2005; Thabet, Stretch, & Vostanis, 2000). Administration by interviewers or interpreters may elicit socially desirable or acquiescent responses (Okawa, 2008; Van de Vijver & Leung, 2011).

3.2.3 | Structural equivalence

Structurally equivalent instruments show identity of underlying dimensions with the original, with comparable factor structure, item loadings, internal and test-retest reliability, and a structure that is meaningful in the target cultural group (AERA et al., 2014; Arnold & Matus, 2000; Fischer & Fontaine, 2011). To this end, ideally measurement invariance is established to ensure latent constructs, and hence scores, are comparable across language groups (Milfont & Fischer, 2010). Measurement invariance is established by first testing for configural invariance, using CFA in each group to determine whether the fit of the factorial configuration is consistent with the original model. If confirmed, multi-group CFA is used to test whether different groups' responses to items and factor structures are comparable, thereby establishing metric invariance. The CFA model may be varied to allow identification of partial measurement invariance by imposing equality constraints on item thresholds, factor loadings or error variances across groups. Significant differences between the constrained and unconstrained models indicate full invariance is not supported (Milfont & Fischer, 2010; Ortuño-Sierra et al., 2015). Finally, scalar invariance – justifying mean score comparisons – is tested by constraining item intercepts to be equal across groups (Milfont & Fischer, 2010).

Thirteen translation studies investigated structural equivalence, including three with immigrant groups. Table 2 shows mean item loadings weighted by sample size (Field, 2001), across 10 studies; three studies did not report loadings. Findings are reviewed in the sequence of languages in alphabetical order, with single-language studies first, followed by studies involving multiple language and immigrant groups.

Principal components analysis (PCA) of an Arabic S-SDQ yielded five factors, but factor loadings were not reported (not included in Table 2 data). A descriptive account indicates that multiple items did not load on their hypothesized factors (Almaqrabi & Shuwail, 2004). CFA of an Arabic P-SDQ, and further unspecified factor analysis, showed all factors were heterogeneous (Thabet et al., 2000).

TABLE 3 Comparative internal consistency (Cronbach alphas) and test-retest reliability of the Strengths and Difficulties Questionnaire (SDQ)

Subscales	P-SDQ			T-SDQ			S-SDQ		
	English, Goodman, 2001	Translations		English, Goodman, 2001	Translations		English, Goodman, 2001	Translations	
		M	Range		M	Range		M	Range
	N = 9,998	Total N = 8,658 ^a		N = 7,313	Total N = 11,351		N = 3,983	Total N = 7,684	
<i>Internal consistency^b</i>									
Total difficulties	0.82	0.75	0.59–0.81	0.87	0.74	0.60–0.84	0.80	0.75	0.57–0.81
Emotional symptoms	0.67	0.63	0.50–0.70	0.78	0.69	0.63–0.76	0.66	0.73	0.59–0.80
Conduct problems	0.63	0.56	0.31–0.63	0.74	0.68	0.62–0.79	0.60	0.56	0.33–0.70
Hyperactivity	0.77	0.74	0.55–0.76	0.88	0.82	0.66–0.88	0.67	0.65	0.55–0.73
Peer problems	0.57	0.35	0.01–0.45	0.70	0.52	0.11–0.74	0.41	0.50	0.28–0.70
Prosocial behaviour	0.65	0.68	0.60–0.70	0.84	0.82	0.77–0.84	0.66	0.69	0.56–0.87
<i>Test–retest reliability^c</i>									
	N = 2,091	Total N = 536		N = 796	Total N = 777		N = 781	Total N = 366	
Total difficulties	0.72	0.84	0.72–0.86	0.80	0.89	0.55–0.90	0.62	0.71	0.62–0.71
Emotional symptoms	0.57	0.80	0.47–0.82	0.65	0.79	0.40–0.80	0.57	0.83	0.70–0.89
Conduct problems	0.64	0.75	0.64–0.75	0.69	0.85	0.50–0.86	0.51	0.61	0.51–0.67
Hyperactivity	0.72	0.84	0.48–0.86	0.82	0.88	0.64–0.89	0.60	0.76	0.61–0.84
Peer problems	0.61	0.75	0.75–0.79	0.72	0.80	0.58–0.81	0.54	0.59	0.45–0.67
Prosocial behaviour	0.61	0.76	0.43–0.78	0.74	0.83	0.50–0.84	0.51	0.75	0.61–0.83

Note: Mean correlations for translation studies are Z-transformed and weighted mean correlations.

^aSee Table 1 for N in individual studies.

^bIncluded internal consistency translation studies: Arabic P-SDQ, Thabet et al., 2000; Chinese P-, T-, S-SDQ, Du et al., 2008; Chinese P-, T-SDQ, Lai et al., 2010; Chinese S-SDQ, Liu et al., 2013; Chinese S-SDQ, Yao et al., 2009; Dari, Pashtu P-, S-SDQ, Panter-Brick et al., 2011; French (Congo) T-SDQ, Kashala et al., 2005; Netherlands Moroccan/Arabic ethnic Turkish ethnic P-, T-SDQ, Mieloo et al., 2014; Netherlands Moroccan/Arabic ethnic, Turkish ethnic, Zwirs et al., 2011; Norway S-SDQ Pakistani ethnic, Other ethnic Richter et al., 2011; Persian/Farsi P-, T-, S-SDQ, Ghanizadeh et al., 2007; Persian/Farsi P-, T-SDQ, Shahrivar et al., 2009; Russian S-, T-SDQ Ruchkin et al., 2007, 2012; Tamil S-SDQ, Lukumar et al., 2008. $k = 15$

^cTest-retest intervals: Goodman, 2001, 4–6 months, Chinese P-, T-SDQ, 6-weeks, Du et al., 2008; Chinese S-SDQ, 2 months, Yao et al., 2009; Chinese P-, T-SDQ, 2–4 weeks, Lai et al., 2010; Chinese S-SDQ, 4 weeks, Liu et al., 2013; Dari, Pashtu, P-, S-SDQ, 1 week, Panter-Brick et al., 2011. $k = 5$. Correlation ranges are primarily based on only two (differing) studies.

TABLE 4 Comparative subscale intercorrelations of the Strengths and Difficulties Questionnaire (SDQ)

Subscales	P-SDQ			T-SDQ			S-SDQ		
	English, Goodman et al., 2010 ^a	Translations ^b		English, Goodman et al., 2010	Translations		English, Goodman et al., 2010	Translations	
	N = 18,222	M	Range	N = 14,263	M	Range	N = 7,678	M	Range
	Total N = 6,199 ^c			Total N = 8,935 ^c			Total N = 10,961 ^c		
Emotion/Conduct	0.51	0.25	0.22–0.36	0.34	0.20	0.12–0.29	0.53	0.36	0.23–0.59
Emotion/Hyperactivity	0.40	0.27	0.21–0.36	0.33	0.27	0.13–0.40	0.48	0.48	0.33–0.64
Emotion/Peer problems	0.71	0.44	0.44–0.44	0.66	0.48	0.25–0.56	0.69	0.44	0.27–0.53
Conduct/Hyperactivity	0.71	0.52	0.46–0.55	0.81	0.68	0.60–0.85	0.85	0.73	0.38–0.92
Conduct/Peer problems	0.58	0.43	0.43–0.43	0.67	0.56	0.29–0.64	0.47	0.36	0.19–0.66
Hyperactivity/Peer prob	0.49	0.33	0.33	0.54	0.40	0.18–0.46	0.38	0.17	0.08–0.50

Note: Means for translation studies are Z-transformed and weighted mean correlations.

^aComparisons are made with Goodman et al. (2010), as Goodman (2001) made only three subscale comparisons.

^bTranslation studies included: Chinese, P-, T-, S-SDQ, Du et al., 2008; Chinese, P-, T-SDQ, Lai et al., 2010; Chinese, S-SDQ, Yao et al., 2009; Netherlands Moroccan/Arabic, Turkish, T-SDQ, Zwirs et al., 2011; Norway, Pakistani ethnic, Other ethnic, S-SDQ, Richter et al., 2011; Persian/Farsi P-, T-SDQ, Shahrivar et al., 2009; Russian, S-, T-SDQ, Ruchkin et al., 2007, 2012; Tamil, S-SDQ, Lukumar et al., 2008. $k = 9$.

^cSee Table 1 for N in individual studies.

CFA of a Chinese S-SDQ showed structural equivalence with a satisfactory fit of the 5-factor model; factor loadings ranged from 0.38 to 0.85 (Yao et al., 2009; loadings were not further reported: not included in Table 2). PCA of Chinese P-, T-, and S-SDQs found inconsistent 5-factor structures across informant groups, with poor support for the conduct and peer problems subscales (Du et al., 2008). Also

with Chinese SDQs, Liu et al. (2013) reported using exploratory factor analysis (EFA) and PCA, finding differing 4-factor structures for the P- and T-SDQs, and a 5-factor structure for the S-SDQ.

Although CFA of the Russian S-SDQ showed acceptable fit to the 5-factor model, Ruchkin, Koposov, and Schwab-Stone (2007) questioned its structural invariance because of lower than expected

TABLE 5 Comparative informant intercorrelations of the Strengths and Difficulties Questionnaire (SDQ)

Informants Subscales	Parent/Teacher			Parent/Self			Teacher/Self		
	English, Goodman, 2001	Translations ^a		English, Goodman, 2001	Translations		English, Goodman, 2001	Translations	
	N = 7,313	M	Range	N = 3,983	M	Range	N = 9,998	M	Range
	Total N = 7,728			Total N = 2,659			Total N = 1,226		
Total difficulties	0.46	0.42	0.20–0.71	0.48	0.54	0.39–0.83	0.33	0.42	0.42
Emotional symptoms	0.27	0.22	0.09–0.56	0.37	0.49	0.42–0.78	0.21	0.18	0.04–0.29
Conduct problems	0.37	0.33	0.16–0.59	0.44	0.45	0.39–0.74	0.30	0.27	0.18–0.34
Hyperactivity	0.48	0.45	0.18–0.67	0.41	0.44	0.37–0.76	0.32	0.32	0.23–0.38
Peer problems	0.37	0.24	0.08–0.29	0.40	0.38	0.36–0.49	0.29	0.32	0.29–0.35
Prosocial behaviour	0.25	0.33	0.12–0.65	0.30	0.45	0.40–0.68	0.23	0.23	0.13–0.31

Note: Means for translation studies are Z-transformed and weighted mean correlations.

^aSee Table 1 for N in individual studies. Translation studies included: Arabic, P/T-SDQ Thabet et al., 2000; Chinese, P/T/S-SDQ Du et al., 2008 (only study that reported TD T/S correlations); Chinese, P/T/S-SDQ Lai et al., 2010; Dari, Pashtu, P/S-SDQ Panter-Brick et al., 2011; Netherlands Moroccan/Arabic, Turkish, P/T-SDQ Mieloo et al., 2014; Persian/Farsi, P/T-SDQ Ghanizadeh et al., 2007; Russian, T/S-SDQ Ruchkin et al., 2012. $k = 7$

TABLE 6 Comparative evidence for convergent validity of the Strengths and Difficulties Questionnaire (SDQ) with the Child Behaviour Checklist (CBCL)

SDQ/CBCL subscales	P-SDQ			T-SDQ ^a		S-SDQ ^b
	English, Goodman & Scott, 1999	Translations		Translations		Chinese
	N = 132	M	Range	M	Range	Yao et al., 2009
	Total N = 1,884			Total N = 2,170		N = 1,135
Total/Total	0.87	0.75	NA	0.71	NA	0.69
Emotional/Internalizing	0.74	0.57	0.49–0.63	0.67	0.52–0.73	0.6
Conduct/Externalizing	0.84	0.46	0.24–0.65	0.79	0.70–0.82	0.61
Hyperactivity/Inattention	0.71	0.54	0.34–0.65	0.42	0.16–0.67	0.65
Peer problems/Social problems	0.59	0.22	0.04–0.51	0.49	0.45–0.53	0.46

Note: Means for translation studies are Z-transformed and weighted mean correlations. Included translation studies: Netherlands, Moroccan/Arabic, Turkish, P-, T-SDQ, Mieloo et al., 2014; Persian/Farsi, P-, T-SDQ, Shahrivar et al., 2009. Only Yao et al. (2009) examined convergent validity with the S-SDQ. $k = 3$.

^aConvergent measure is the CBCL Teacher Report Form.

^bConvergent measure is the CBCL Youth Self Report.

factor loadings and internal consistency, and high factor intercorrelations. The Russian T-SDQ showed similar but somewhat stronger structural properties than the Russian S-SDQ (Ruchkin, Koposov, Vermeiren, & Schwab-Stone, 2012).

PCA of the French T-SDQ, used in the DR of Congo, identified five unnamed factors that, from factor loadings may be interpreted as: positive construal, hyperactivity, emotional symptoms, conduct problems, and externalization (Kashala, Elgen, Sommerfelt, & Tylleskar, 2005). No peer problem factor was identified and multiple items loaded or cross-loaded on non-designated factors.

PCA of the Persian/Farsi P-, T-, and S-SDQ showed that conduct and peer problem items did not load on hypothesized factors and a 3-factor structure (externalization, internalization and positive construal) was more readily interpreted than a 5-factor model (Ghanizadeh, Izadpanah, & Abdollahi, 2007; factor loadings not reported: not included in Table 2).

In the Netherlands, individual and multi-group CFA of a Dutch T-SDQ for four ethnic groups (including Moroccan and Turkish), established configural, metric and scalar invariance, with acceptable fit for the 5-factor model across groups (Zwirs et al., 2011). Also in the Netherlands PCA with a forced 5-factor solution of P- and T-SDQs

with Dutch, Moroccan and Turkish children, found inconsistent factor structure between groups (Mieloo et al., 2014). Dutch parents' SDQ items loaded on predicted factors, but for Moroccan and Turkish parents' SDQs three and two factors, respectively, yielded no identifiable structure. Dutch teachers' SDQ showed factor equivalence with similar 5-factor structures for ethnic and Dutch groups (Mieloo et al., 2014).

In Norway, individual-group CFA of the S-SDQ showed the 5-factor model provided a good fit for Norwegian youth, and an acceptable fit for Pakistani and Other (Iranian, Turkish and Somali) youth, indicating configural invariance. However, multi-group CFA did not support metric invariance as thresholds for response categories and factor loadings differed significantly between Norwegian and refugee and immigrant groups (Richter, Sagatun, Heyerdahl, Oppedal, & Røysamb, 2011).

Multi-group CFA of the S-SDQ across seven countries, including Turkey and Serbia, found no satisfactory fit for the 5-factor model (Stevanovic et al., 2014). Without configural invariance, metric and scalar invariance across countries could not be tested. Subsequent EFA identified three common factors (prosocial, emotional symptoms, and conduct problems) based on only 10 replicable items across all countries (Stevanovic et al., 2014).

TABLE 7 Comparative evidence for the predictive validity of the Strengths and Difficulties Questionnaire (SDQ) (in language order)

Language, Authors Criterion measures	Scale	P-SDQ ^a			T-SDQ ^a			S-SDQ ^a		
		Sensitivity/ range	Specificity/ range	AUC/range (Odds ratios)	Sensitivity/ range	Specificity/ range	AUC/range (Odds ratios)	Sensitivity/ range	Specificity/ range	AUC/range (Odds ratios)
Reference studies										
English, Goodman et al., 1998	TD	—	—	—	—	—	—	—	—	0.82
Clinic/Community	Subscales	—	—	—	—	—	—	—	—	0.65–0.77
English, Goodman, 2001	TD	47%	94%	15.2	43%	95%	13.5	23%	94%	6.6
DAWBA DSM-IV diagnosis	Subscales	25–74%	91–96%	3.4–32.3	29–68%	91–95%	5.2–33.4	13–40%	89–96%	1.7–9.7
Translation studies										
Arabic, Almagrabi & Shuwail, 2004	TD	—	—	—	—	—	—	61%	81%	0.77
Clinic/community	Subscales	—	—	—	—	—	—	—	—	0.80–0.89
ICD-10 diagnosis	Algorithm	—	—	—	—	—	—	72%	55%	—
Arabic, Alyahri & Goodman, 2006	TD	—	—	0.81 (.74–.87)	—	—	0.76 (.68–.83)	—	—	—
Clinic/Community	Subscales	—	—	0.70–0.97	—	—	0.65–0.97	—	—	—
DAWBA DSM-IV diagnosis	E, C, H	—	—	0.76–0.89	—	—	0.72–0.89	—	—	—
Chinese, Du et al., 2008	TD	—	—	0.69 (.71–.83)	—	—	0.91 (.87–.95)	—	—	0.72 (.64–.80)
ADHD/controls	Subscales	—	—	0.39–0.77	—	—	0.60–0.90	—	—	0.61–0.70
Chinese, Lai et al., 2010	TD	—	—	0.84	—	—	0.78	—	—	—
Clinic/community	Subscales	—	—	0.60–0.89	—	—	0.69–0.86	—	—	—
Persian/Farsi, Ghanizadeh et al., 2007	TD	90%	67%	—	—	—	—	—	—	—
K-SADS Diagnosis										
Persian/Farsi, Shahrivar et al., 2009	TD	61%	85%	0.81	56%	31%	0.57	—	—	—
DSM-IV Diagnosis	E, C, H	59–74%	40–98%	0.50–0.97	62–63%	35–79%	0.45–0.76	—	—	—
Russian, Goodman et al., 2005	M-IA	PPV 35%	NPV 86%	—	—	—	—	—	—	—
DAWBA ICD-10 diag.										
Tamil, Lukumar et al., 2008	TD	—	—	—	—	—	—	85.5%	67.2%	0.87 (.79–.93)
Clinic/community										
Urdu, Samad et al., 2005	TD	69%	71%	.77 (.70–.83)	—	—	—	—	—	—
Psychiatric/paediatric	Subscales	59–64%	66–79%	0.73–0.82	—	—	—	—	—	—
ICD-10 Diagnosis/S-SDQ algorithm	E, C, H, Any	33–83%	—	—	—	—	—	—	—	—

Note: Italicized figures represent odds ratios. Figures in parentheses represent 95% confidence intervals. Any = any disorder, AUC = area under the curve; TD = total difficulties, C = conduct disorder, E = emotional disorder, H = hyperkinetic disorder, M-IA = multi-informant algorithm; NPV = negative predictive value; PPV = positive predictive value.

^aSee Table 1 for N values.

TABLE 8 Comparative means and cutoff scores for the Strengths and Difficulties Questionnaire (SDQ)

	P-SDQ				T-SDQ				S-SDQ			
	UK website ^a		Translations		UK website		Translations		UK website		Translations	
	Cutoff	M (SD)	Cut range	M (SD) [ES] ^b	Cutoff	M (SD)	Cut range	M (SD) [ES]	Cutoff	M (SD)	Cutoff ^c	M (SD) [ES]
	N = 10,298		Total N = 7,447 ^{d,e}		N = 8,208		Total N = 10,637 ^e		N = 4,228		Total N = 5,383 ^e	
Total difficulties	17	8.4 (5.8)	15–19	10.9 (5.7) [0.43]	16	6.6 (6.0)	13–21	8.9 (5.5) [0.40]	20	10.3 (5.2)	18	11.0 (4.7) [0.14]
Emotional symptoms	5	1.9 (2.0)	4–5	2.2 (1.9) [0.15]	6	1.4 (1.9)	3–6	1.8 (1.8) [0.21]	7	2.8 (2.1)	6	2.6 (2.1) [0.10]
Conduct problems	4	1.6 (1.7)	4	1.9 (1.6) [0.18]	4	0.9 (1.6)	3–5	1.4 (1.7) [0.30]	5	2.2 (1.7)	5	2.7 (1.6) [0.30]
Hyperactivity	7	3.5 (2.6)	6–8	4.2 (2.3) [0.28]	7	2.9 (2.8)	7–9	3.4 (2.6) [0.19]	7	3.8 (2.2)	7	3.7 (2.0) [0.1]
Peer problems	4	1.5 (1.7)	4–6	2.5 (1.7) [0.59]	5	1.4 (1.8)	3–6	2.2 (1.8) [0.44]	6	1.5 (1.4)	6	2.6 (1.7) [0.71]
Prosocial behaviour	0–4	8.6 (1.6)	0–4/5	7.1 (1.9) [0.90]	0–4	7.2 (2.4)	0–3/4	6.8 (2.4) [0.17]	0–4	8.0 (1.7)	0–4	6.7 (2.2) [0.66]

Note: Means for translation studies are weighted means (M values); standard deviations (SD values) are calculated from pooled weighted variances. Included translation studies reporting means: Arabic, S-SDQ, Almqarimi & Shuwail, 2004; Chinese, P-, T-, S-SDQ, Du et al., 2008; Chinese, P-, T-SDQ, Lai et al., 2010; Chinese, S-SDQ, Yao et al., 2009; Dari, Pashtu P, S-SDQ, Panter-Brick et al., 2011; French (Congo), T-SDQ, Kashala et al., 2005; Netherlands Moroccan/Arabic ethnic, Turkish ethnic, P-, T-SDQ, Mieloo et al., 2014; Netherlands Moroccan/Arabic ethnic, Turkish ethnic, T-SDQ, Zwirs et al., 2011; Persian/Farsi, P-, T-SDQ, Shahrivar et al., 2009; Russian, S-, T-SDQ, Ruchkin et al., 2007, 2012; Serbian, Turkish, S-SDQ, Stevanovic et al., 2014; Urdu, P-SDQ, Samad et al., 2005. $k = 12$.

^ayouthmind (2013).

^bEffect sizes (ES) for differences from UK means.

^cOnly Du et al. (2008) reported S-SDQ cutoffs.

^dSee Table 1 for N in individual studies.

^eTotal N for total difficulties: P-SDQ 7,951, T-SDQ 8,965, S-SDQ 3,455.

Ten of the 13 factor analytic studies identified a positive construal factor, whereby up to five of the positively-worded “problem” items loaded or cross-loaded on the prosocial factor in PCA or EFA studies, or showed low loadings in CFA studies. Additional items that repeatedly failed to load on predicted factors included: *better with adults* (five studies); *temper and bullied* (four); and *solitary and restless* (three).

3.2.4 | Equivalence in reliability

Fifteen translation studies reported internal consistency, including four that did not conduct factor analysis (Lai et al., 2010; Lukumar et al., 2008; Panter-Brick et al., 2011; Shahrivar, Tehrani-Doost, Pakbaz, Rezaie, & Ahmadi, 2009). Notably, studies that found structural inequivalence reported internal consistency for the standard subscales, not for the identified factors (e.g. Du et al., 2008). To obtain mean Cronbach alphas for translation studies (Table 3), correlations were Z-transformed, weighted by sample size, and then back-transformed to a correlation coefficient (Field, 2001). Similar calculations were conducted for correlations in Tables 4–6.

Across informants, Cronbach alphas for total difficulties tended to be lower than in the United Kingdom. The hyperactivity subscale generally showed the highest alphas, and peer problems the lowest. Where P- and T-SDQs were both used, reliabilities were consistently higher for the T-SDQ.

For five studies reporting test–retest reliabilities, retest intervals ranged from 2 to 8 weeks, compared to 4–6 months for the original SDQ (Table 3). This difference in interval may explain why translated total difficulties test–retest correlations tended to be higher than the UK SDQ. As correlation ranges show (Table 3), subscale stability varied considerably within and across studies, but correlations appeared to be low to acceptable.

3.2.5 | Item and scalar equivalence

While CFA may identify structural inequivalence and group differences in item endorsement thresholds, item response theory methods may be used to identify differential item functioning (DIF), an alternative method of testing measurement invariance, which at times provides different results (Meade & Lautenschlager, 2004). Sireci (2011) argues that differential item analysis should follow investigation of structural inequivalence. DIF occurs when groups matched on a characteristic systematically provide different responses to an item when similar responses are expected (AERA et al., 2014; Sireci, 2011). DIF may reflect construct-irrelevant variance, such as inaccurate translation, or cultural influences, or may reflect “true” differences between groups. The reasons for DIF require investigation to avoid scalar inequivalence. Tests that have scalar equivalence have the same units of measurement, hence item and total mean scores may justifiably be compared across cultures (AERA et al., 2014; He & van de Vijver, 2012).

Richter et al. (2011) concluded that mean scores could not be compared between Norwegian and ethnic groups because significant differences were found in S-SDQ item endorsement thresholds, attributing the differences to cultural variations in response styles, or in conceptions of problematic behaviour. No reviewed studies analysed

DIF, although Mieloo et al. (2014) acknowledged the need to investigate DIF, which may have influenced the non-invariant factor structures across Netherlands ethnic groups.

3.2.6 | Equivalence in subscale and informant intercorrelations

In eight of nine translation studies reporting subscale intercorrelations, conduct and hyperactivity subscales showed the highest correlations, consistent with Goodman et al.'s (2010) second-order externalizing factor (Table 4). Emotional subscale intercorrelations generally showed appropriate differentiation from the externalizing subscales, but the emotional and peer problems subscales did not show high correlations consistent with an internalizing factor.

Seven translation studies reported informant intercorrelations (Table 5). Parent/teacher correlations were consistently lower than in the United Kingdom, with all subscales showing almost an absence of correlation at the lower end of the range, as found particularly by Mieloo et al. (2014), with Moroccan and Turkish children. Parent/self-correlations were similar to or somewhat higher than UK values, while teacher/self-correlations were comparable, with the exception of the emotional subscale, which was particularly low for Ruchkin et al. (2012; $r = 0.04$). Ghanizadeh et al. (2007) showed notably high parent/teacher and parent/self-agreement with respective ranges of $r = 0.28$ – 0.71 , and 0.49 – 0.83 , for peer problems and total difficulties, respectively.

3.2.7 | Convergent equivalence

Cross-cultural convergent equivalence is shown when the original and translated SDQs show a similar pattern of relations with instruments assessing similar constructs (Van de Vijver & Leung, 2011). Of six translation studies that examined convergent validity evidence, three investigated convergence with translated CBCLs (Table 6), with two reporting validated CBCL translations (Shahrivar et al., 2009; Yao et al., 2009). SDQ/CBCL correlations were generally lower but showed a similar pattern to the United Kingdom (Goodman & Scott, 1999). SDQ peer problems/CBCL social problems correlations tended to be lowest in all studies, possibly due to the peer problems subscale's poor factor structure. In the Netherlands convergence was particularly poor for Moroccan children on the P-SDQ/CBCL conduct/externalizing subscales ($r = 0.24$), and the T-SDQ/CBCL hyperactivity/inattention subscales ($r = 0.16$) (Mieloo et al., 2014).

Correlations with other convergent validity measures (not shown in Table 6) were as follows: the Chinese P-SDQ (Du et al., 2008) and validated Chinese Conner's Parent Symptom Questionnaire correlated $r = 0.63$ (Su, Li, Huang, Luo, & Zhang, 2001); the Dari and Pashtu S-SDQ (Panter-Brick et al., 2011) correlated with the validated Depression Self Rating Scale $r = 0.5$ (Birleson, Hudson, Buchanan, & Wolff, 1987) and Child Revised Impact of Events Scale $r = 0.42$ (Perrin, Meiser-Stedman, & Smith, 2005). The Russian S-SDQ hyperactivity and conduct scales (Ruchkin et al., 2012), correlated $r = 0.58$ – 0.78 with the Teacher Hyperactivity ADHD, and Swanson, Nolan, and Pelham (SNAP-IV) Oppositional Defiant Disorder and Conduct Disorder (Swanson, 1992) measures (validation not reported).

3.2.8 | Predictive equivalence

Predictive, criterion-related, or discriminant equivalence is shown if translated tests predict caseness as accurately on a designated "gold-standard" diagnostic instrument as the original (Arnold & Matus, 2000). Evidence for predictive validity is generally tested through receiver operating characteristic (ROC) curve analysis, which plots sensitivity (true positives) against 1-specificity (false positives), providing a trade-off between the two. The AUC combines information on sensitivity, and specificity, representing the probability that cases and non-cases will be correctly differentiated across each cut-point. Balanced sensitivity and specificity ≥ 0.80 should be shown at optimal cutoff scores. An AUC of 1.0 represents perfect accuracy, ≥ 0.8 moderate to high accuracy, and 0.5 denotes chance detection (Akobeng, 2007; Lalkhen & McCluskey, 2008).

Equivalence with the original SDQ is not readily established as evident in Table 7. Goodman (2001) reported sensitivity, specificity and odds ratios, while none of the nine translation studies that investigated criterion equivalence reported odds ratios. Moreover, translation studies differed in whether AUCs, sensitivity and specificity were all reported and whether for TDSs only or for subscales as well. Consequently predictive evidence could not meaningfully be summarized across studies; Table 7 shows total difficulties and subscale ranges for available ROC values for each translation study.

We review studies in the sequence of six studies that examined the SDQ's capacity to discriminate between clinical and community or normal control groups, including three that also predicted diagnoses, and three studies that examined predictive accuracy for diagnoses only.

The Hong Kong Chinese P- and T-SDQ total difficulties and subscale scores discriminated well between community and clinic samples with AUCs ≥ 0.80 . The peer problems subscale for both informants and the T-SDQ emotion subscale provided exceptions, showing AUCs ≤ 0.70 (Lai et al., 2010). Sensitivity and specificity were not reported.

The Tamil S-SDQ TDS showed moderate accuracy in differentiating clinic youth from community controls, with a range of cut-points showing that 13 provided the most balanced sensitivity and specificity of 77.4 and 76.6, respectively (Lukumar et al., 2008). An AUC (0.87) was reported for a 12 cutoff but not for other cutoffs. Subscale ROC values were not reported.

A high AUC of 0.91 for the mainland Chinese T-SDQ TDS better differentiated ADHD clinic outpatients from community controls than P- and S-SDQ TDSs. The hyperactivity subscale better identified ADHD cases than other subscales (Du et al., 2008).

In Taiwan psychiatrist-diagnosed DSM-IV ADHD cases were differentiated from school-based controls by significantly higher mean scores on the Chinese P-, T-, and S-SDQ emotional and conduct subscales, and lower S-SDQ hyperactivity scores (Liu et al., 2013). The latter was attributed to ADHD youth under-estimating their own hyperactive behaviour (data not shown due to inconsistent factors).

Arabic S-SDQ scores differentiated clinic from community youth, with satisfactory AUCs ranging from 0.77 (total difficulties) to 0.89 (conduct). However, TDS sensitivity was a low 61% and specificity a satisfactory 81%. In the clinic the S-SDQ algorithm's sensitivity for any clinician-assigned ICD-10 (WHO, 1992) diagnosis increased to 72% but specificity decreased to 55% (Almaqrabi & Shuwail, 2004), which would yield 45% false negatives.

Yemeni clinic and community children were moderately to well-differentiated by Arabic P-SDQ scores: AUCs ranged from 0.70 (peer problems) to 0.97 (hyperactivity; Alyahri & Goodman, 2006). T-SDQ AUCs were similar. In the clinic AUCs for predicting DAWBA-based DSM-IV diagnoses from P- and T-SDQs ranged from 0.76 and 0.72, respectively, for emotional disorders, to 0.89 for both informants for conduct disorders (Alyahri & Goodman, 2006). Sensitivity and specificity were not reported.

The Urdu P-SDQ total difficulties and subscale scores differentiated psychiatric from paediatric patients, with a moderate total difficulties AUC score of 0.77, balanced sensitivity (0.69%) and specificity (0.71%), and significantly higher mean TDSs. In the psychiatric group the P-SDQ algorithm rather poorly predicted psychiatrist-assigned ICD-10 disorders with percentages correctly identified ranging from 33% of hyperkinetic cases, 56% of any disorder, to 83% (of six cases) of conduct disorders (Samad et al., 2005).

In Iran the Persian/Farsi P-SDQ TDS identified any DSM-IV diagnosis on the Kiddie Schedule for Affective Disorders and Schizophrenia – Present and Lifetime Version (Ghanizadeh, Mohammadi, & Yazdanshenas, 2006), with 90% sensitivity and 67% specificity (Ghanizadeh et al., 2007).

The Persian/Farsi P-SDQ TDS differentiated children with psychiatrist-assigned DSM-IV diagnoses from non-diagnosed children, with a satisfactory 0.81 AUC, 61% sensitivity and 85% specificity. P-SDQ ROC values were highest for detection of hyperkinetic disorders and lowest for emotional disorders. T-SDQ ROC values tended to be lower than for the P-SDQ (Shahrivar et al., 2009).

A Russian SDQ M-IA yielded low 35% positive predictive value (PPV) and 86% negative predictive value (NPV) for DAWBA-based ICD-10 diagnoses of SDQ screen-positive and screen-negative school children (Goodman, Slobodskaya, & Knyazev, 2005). PPVs and NPVs represent the proportion of true positives and true negatives, respectively, but unlike sensitivity and specificity, depend on the population prevalence of disorder (Lalkhen & McCluskey, 2008). The low PPV would yield >60% false positives.

3.2.9 | Normative equivalence

Normative data enable comparisons to be made between test takers' performance and others of a similar background (AERA et al., 2014). Findings of significant differences in mean scores between groups on the original and translated tests require investigation to determine whether these reflect valid differences or cultural influences on measurement (Van de Vijver & Poortinga, 2005). New norms and cutoffs may need to be developed to establish normative equivalence (Arnold & Matus, 2000).

Norms, means and/or cutoff scores for the original and seven translated SDQs are available at youthinmind (2013), but none are refugee-relevant. Table 8 shows UK mean scores (youthinmind, 2013) compared with weighted means averaged across 12 reviewed studies. Means for translation studies were weighted for respective studies' sample sizes; standard deviations were calculated from pooled variances weighted for sample sizes and number of studies minus degrees of freedom, then converted back to standard deviations (Howell, 2002). For seven studies that provided means for boys and girls data

were similarly weighted and pooled to obtain total samples. Table 8 also shows effect sizes (ESs) in brackets for differences from UK means. ESs were calculated using the formula $M1 - M2/\text{mean standard deviation (SD) of } M1 \text{ and } M2$ (Sullivan & Feinn, 2012). Cohen deemed ESs as: $d = 0.2$ small, $d = 0.50$ medium, and $d \geq 0.80$ large (Sullivan & Feinn, 2012).

Of the 12 studies, six established normative equivalence by developing local cutoffs for a total of 13 groups (informants and ethnic groups): The moderate to large ESs for predominantly higher than UK means demonstrate the need for local cutoffs. In particular, across the majority of translation studies the peer problems and prosocial subscales consistently showed moderate to large ESs for higher and lower means, respectively, than in the United Kingdom.

The importance of local cutoffs is further highlighted by two additional studies that reported percentages scoring above UK cutoffs (nominally 10% in the United Kingdom). On the Russian M-IA 19% scored above the UK cutoff (Goodman et al., 2005), as did 70.7% of three-year-olds on the Arabic P-SDQ emotional subscale, reducing to 34.9% across all age groups (Thabet et al., 2000). The absence of local cutoffs leaves unanswered whether these are true differences in disorder or cultural effects.

3.2.10 | Sensitivity to change following interventions with ethnic groups

No SDQ outcome studies were located that were conducted in refugee-source countries. Seven studies were undertaken in English-speaking countries with immigrant and refugee groups. Related studies are reported together.

Three studies used the T-SDQ to evaluate a London school-based refugee mental health service for children referred by teachers (Dura-Vila, Klasen, Makatini, Rahimi, & Hodes, 2013; Hodes, 2002; OShea, Hodes, Down, & Bramley, 2000). Dura-Vila et al. (2013) also administered the P-SDQ, for which parents were provided with interpreters if necessary. However, some of the languages spoken by parents (Kurdish and Somali) have no validated translations on the SDQ website. OShea et al. (2000) showed no significant improvement in T-SDQ scores, but with a larger sample, Dura-Vila et al. (2013) found significant reductions in P- and T-SDQ total difficulties and subscale scores.

In Canada creative expression sand-play workshops aimed to reduce emotional and behavioural problems of immigrant and refugee pre-schoolers (Rousseau, Benoit, Lacroix, & Gauthier, 2009). Bilingual telephone interviewers administered the P-SDQ in (unspecified) native languages. For the whole intervention group P-SDQ total difficulties, emotional and peer problem scores decreased significantly, but not T-SDQ scores. For children whose families had experienced pre-migration adversity the T-SDQ total difficulties and the P- and T-SDQ impact scale also decreased significantly. Control group T-SDQ emotional and conduct scores increased significantly, attributed to the lack of intervention (Rousseau et al., 2009).

Three studies evaluated classroom-based drama therapy workshops for refugee and immigrant youth in Canada (Rousseau, Armand, Laurin-Lamothe, Gauthier, & Saboundjian, 2012; Rousseau et al., 2007; Rousseau et al., 2014). Students completed S-SDQs in French, English, or students' (unspecified) native languages, supported by multilingual

project staff (Rousseau et al., 2007); with interviewers (Rousseau et al., 2012); or requested translated versions (Rousseau et al., 2014). The three studies found no significant decreases in S- and T-SDQ TDSs (viewed as secondary global measures by Rousseau et al., 2014). However, S-SDQ impact or impairment ratings (viewed as primary outcomes) decreased significantly for Rousseau et al. (2007) and Rousseau et al. (2012), and for first-generation but not second-generation youth (Rousseau et al., 2014). Control groups' S-SDQ impact and T-SDQ TDSs increased significantly (Rousseau et al., 2007; Rousseau et al., 2014), again possibly due to non-intervention.

4 | DISCUSSION

The purpose of this article was to evaluate whether 11 refugee-relevant translations on the SDQ website (youthinmind, 2015) demonstrated cultural equivalence with the original SDQ. As also found by Gadeberg and Norredam (2016), there is a dearth of validation studies of screening instruments with young refugees. However, in the absence of studies validating translated SDQs with refugee children and adolescents, our review focused on translations validated in refugee-source countries, or with immigrant children from those countries. As well as meeting a recognized need for examination of the validity of cross-cultural screening tools (Gadeberg & Norredam, 2016), this review is the first to examine the possible effects of translation on the psychometric properties of translated SDQs.

In evaluating cultural equivalence, the moderately satisfactory psychometric properties of the original SDQ need to be borne in mind. Evidence for content validity was based on the RPQ, investigation of response process was somewhat informal, and structural validity evidence showed that a second-order 3-factor model provided a better fit (Goodman et al., 2010) than the original 5-factor model (Goodman, 2001). The 5-factor model's internal consistency was moderate to high, with the exception of the peer problems factor. Inter-informant correlations generally exceeded Achenbach et al.'s (1987) meta-analytic mean, and the SDQ and CBCL showed moderate to high convergence. Evidence for clinical predictive validity was equivocal, with low sensitivity and high specificity, although sensitivity improved with use of the M-IA (Goodman, 2001; Goodman et al., 2003).

4.1 | Cultural equivalence of translated SDQs

The remainder of the discussion considers evidence for the cultural equivalence of translated SDQs.

No reviewed study reported on possible changes in item connotations resulting from translation, or on challenges in achieving conceptual and linguistic equivalence. Examples of such changes are described by Tesfaye, Hanlon, Wondimagegn, and Alem (2009) in reporting on Ethiopian women's misunderstandings of Kessler-10 items. However, altered SDQ connotations are suggested, for example, by the failure of the item *better with adults*, to load in five translation studies. Although scored as a problem item in the UK SDQ, better relationships with adults were viewed as a positive attribute by Australian Indigenous parents (Williamson et al., 2014), and as

evidence of maturity by Thai parents (Woerner, Nuanmanee, Becker, Wongpiromsarn, & Mongkol, 2011).

Twelve of the 20 validation studies acknowledged possible influences of differing cultural values when findings did not match UK findings. Higher mean scores on the Chinese than the UK P-SDQ were attributed to more "stringent" expectations of obedience in the competitive Hong Kong culture (Lai et al., 2010). Yao et al. (2009) ascribed high Chinese S-SDQ peer problem scores to Chinese collectivist values of greater social sensitivity, an argument that could explain high peer problems scores found in the majority of studies. Thabet et al. (2000) argued that three-year-olds' high scores on the Arabic P-SDQ emotional subscale may be normative for toddlers in Arab culture, which endorses the protective proximity of the mother. However, supporting evidence or citations for these speculations were lacking.

No reviewed study discussed whether response scales retained similar meaning in translation. The fine gradations of emotions required by Likert-type scales may be unfamiliar in non-Western cultures (Kinzie et al., 1982). Changes in response scale connotations could alter item endorsement thresholds, potentially influencing factor structure and mean scores (AERA et al., 2014; Van de Vijver & Leung, 2011).

The method of administering the SDQ was largely unexamined, although Thabet et al. (2000) acknowledged potential interviewer influence on Arabic P-SDQ responses. Mode-of-administration research shows that, compared to self-completed measures, personally interviewed respondents are less likely to disclose sensitive topics such as mental illness (Bowling, 2005). Some findings of high parent/youth intercorrelations raise questions about the independence of responses to P- and S-SDQs sent home for completion.

Factor analysis of 13 translated SDQs yielded mixed results. Seven studies conducted CFA, the recommended method to confirm a theoretical structure and measurement invariance (Fischer & Fontaine, 2011; Milfont & Fischer, 2010). Four of these studies were single language-group studies that were therefore only able to test for configural invariance: only Yao et al. (2009) clearly confirmed the 5-factor model. The remaining three CFA studies made inter-country (Stevanovic et al., 2014) or ethnic-group comparisons (Richter et al., 2011; Zwirs et al., 2011): only the latter demonstrated full measurement invariance.

Professed 5-factor solutions by six PCA studies were not well-supported by the evidence, as multiple items cross-loaded, hypothesized factors failed to emerge, factors were uninterpretable, or factor structures were non-invariant across informants or groups. The peer problems factor in particular contributed to translated SDQs' generally poor structural equivalence, with no items loading on this factor in five studies and low or cross-loadings in another five studies.

Other inter-country comparisons of the S-SDQ, such as made by Stevanovic et al. (2014), also found poor fit for a 5-factor model, with a 3-factor model providing a better but non-invariant fit across five European countries (Germany, Cyprus, England, Sweden, Italy; Essau et al., 2012), and partial measurement invariance with DIF for 11 items across Spain, England, Ireland, Germany and France (Ortuño-Sierra et al., 2015). Ortuño-Sierra et al. (2015) suggested that greater cultural differences may explain greater disparities in factor structure found by Stevanovic et al. (2014), who compared Turkey, Serbia, Bulgaria, Croatia, India, Indonesia, and Nigeria.

In 10 translation studies positively-worded items from the difficulties subscales loaded on a positive construal factor, which supports the view that positively- and negatively-worded items may not measure the same construct, potentially having a deleterious effect on factor structure and internal consistency (Barnette, 2000). Essau et al. (2012) found that removal of reverse-worded items improved fit across European countries.

The structural validity evidence from the translation studies yielded no consistent alternative factor model. Given the structural differences across languages, translated subscales appear to measure different constructs to the original SDQ. To avoid obtaining distorted scores, it may be advisable to adopt Mieloo et al. (2014) and Richter et al.'s (2011) recommendation to use total difficulties rather than subscales scores, thereby providing an overall picture of difficulties. A follow-up clinical interview would aim to identify the specific nature of those difficulties.

The degree of variation found in informant intercorrelations across languages appears to accord with other findings of cross-cultural variations in informant agreement (Grigorenko, Geiser, Slobodskaya, & Francis, 2010; Rousseau & Drapeau, 1998). Lower parent/teacher intercorrelations than in the United Kingdom found in five groups, suggest that the SDQ is not measuring the same construct in each informant group (Arnold & Matus, 2000). T-SDQ factor structures for ethnic groups in settlement countries tended to be more consistent with the UK SDQ (Mieloo et al., 2014; Zwirs et al., 2011), whereas P-SDQ structures differed markedly, which suggests that teachers' ratings of minority children may be more consistent with the original (Western) norms while parents' ratings may be influenced by their culture of origin (Mieloo et al., 2014).

Translated SDQs showed convergence with translated CBCLs and other measures of related constructs, with a similar pattern, if somewhat lower correlations than in the United Kingdom. Exceptions were particularly low SDQ/CBCL subscale correlations for Moroccan children in the Netherlands. Whether this is attributable to translation of the SDQ or CBCL is unknown, however, the largely unidentifiable P-SDQ factor structure for this group may be a contributing factor.

Different indices of predictive accuracy used by studies of the original and translated SDQs complicate comparisons and conclusions regarding predictive equivalence. Based on AUC values, differentiation of clinical and control groups and prediction of diagnoses by translated SDQs generally showed moderate to high accuracy. Total difficulties AUCs predominantly ranged between 0.7 to ≥ 0.8 , with some high and low outliers in the Chinese and Persian/Farsi T-SDQs, respectively. Subscale AUCs were consistently highest for the hyperactivity and conduct subscales, lower for emotional symptoms – respectively representing more and less observable behaviours (Lai et al., 2010) – and lowest for peer problems and prosocial behaviour, which predict any disorder (Goodman, 2001). The poor structural validity of the latter two subscales may explain their low AUCs.

Questions regarding predictive accuracy arise from studies that reported only AUCs, as studies that also reported sensitivity and specificity showed that moderate AUCs may be associated with low sensitivity, as demonstrated by the Arabic S-SDQ (Almaqrabi & Shuwall, 2004) and Urdu P-SDQ (Samad et al., 2005). Both found AUCs of 0.77 for TDSs but low sensitivities of 61% and 69%, respectively, for differentiating clinic from community cases.

Compared to the UK SDQ, translation studies generally showed higher, though still low sensitivity, and somewhat lower, though still high specificity. Lower false positives would lessen the need for further assessment by under-resourced clinical services and reduce unwarranted labelling of children (Goodman et al., 2003), but increased false negatives would elevate the risk of not detecting true cases.

The SDQ's brevity and its cost-free availability in many languages are cited as advantages for developing countries with few child mental health professionals (Almaqrami & Shuwail, 2004; Lukumar et al., 2008; Panter-Brick et al., 2011); instrument brevity may also be of benefit for screening refugee children (Gadeberg & Norredam, 2016). However, the subscales' very brevity and narrow range of behaviours sampled risk construct under-representation, which can impair predictive accuracy (Van de Vijver & Poortinga, 2005) and may explain the SDQ's relative lack of sensitivity in detecting disorders. Ruchkin et al. (2007) questioned whether the SDQ item pool was adequate for devising psychometrically sound scales, arguing that the Russian S-SDQ's factor structure made it unsuitable for its recommended use in clinical assessment.

In part because of varying measures of predictive accuracy used, findings regarding the predictive equivalence of translated SDQs are equivocal. It is difficult to reach conclusions other than that both translated and the original SDQs are likely to yield relatively high false positives.

As observed by Goodman et al. (2012), use of UK norms and comparison of mean scores across cultures may be invalid, potentially leading to over- or under-estimation of disorders and incorrect clinical decisions (AERA et al., 2014; He & van de Vijver, 2012). In the 12 translation studies reporting means, the importance of establishing local cutoffs is evident in the degree to which means differ from UK scores.

Normative data for the reviewed translated SDQs are not available on the SDQ website, but their ready availability would improve valid score interpretations, as would links to relevant validation studies. However, local norms may need to be developed for refugee and immigrant children as norms from respondents' country of origin may not apply in settlement countries (Plitas, Tucker, Kritikos, Walters, & Bardenhagen, 2009).

Of seven outcome studies with refugee and immigrant groups, only two showed statistically significant effects on SDQ TDSs. Three Canadian studies that found significant improvements in S-SDQ impact but not TDSs, overlooked the lack of change in symptom scores by designating impact scores as primary and TDSs as secondary (Rousseau et al., 2014).

The lack of convincing overall findings with ethnic groups leaves unanswered whether the SDQ is insensitive to change with these groups, or whether the programmes were ineffective. Conclusions are confounded by a lack of control groups, poor retention, possibly unvalidated translations, and involvement by interpreters, interviewers or support staff. The narrow range of SDQ outcome studies with ethnic groups further limits conclusions, as only three programmes were evaluated by the seven studies. Given these equivocal findings, and as translated SDQs have predominantly shown structural inequivalence, the SDQ should be used with caution in evaluating cross-cultural mental health interventions (Stevanovic et al., 2014).

Nine outcome studies with English-speaking groups have generally demonstrated the SDQ's sensitivity to change (details available from authors). However, problem-specific measures tended to show larger ESs than the broad-based SDQ (Ford, Hutchings, Bywater, Goodman, & Goodman, 2009; Lee et al., 2005).

4.2 | Implications

Resettled refugee children and their families have been exposed to systematic state-sanctioned violence, persecution, loss, and displacement. The psychological sequelae of these cumulative traumas, combined with resettlement stressors may contribute to complex developmental and mental health needs (Bronstein & Montgomery, 2011; Henley & Robinson, 2011; Kaplan, 2013). Further discussion of these needs is beyond the scope of this article, as is the most effective method for their identification. Suffice to say that a method is required that is sufficiently sensitive to detect the sequelae of trauma. In the absence of item content that would identify these sequelae the SDQ lacks this sensitivity and risks not detecting trauma-related symptoms, beyond perhaps elevated emotional subscale scores (Goodman et al., 2003). No reviewed studies reported association of SDQ scores with trauma assessment scales.

Although apparently not validated with children identified as refugees, the SDQ has been used in detention and asylum-seeker centres, where proportions scoring in the abnormal range have ranged from very high to normative (Lorek et al., 2009; Wiegersma, Stellinga-Boelen, & Reijneveld, 2011). Multiple factors may influence refugees' scores, including conceptually inequivalent translation, interviewer or interpreter administration effects, normative inequivalence, and importantly, situational demand characteristics (Whitehouse, Orne, & Dinges, 2002). Refugee and asylum-seeker-parents and adolescents may deny or over-state mental health problems to improve the likelihood of resettlement and service entitlements (Loutan, Bollini, Pampallona, Haan, & Gariazzo, 1999).

With refugee children, as with any child, the SDQ should not be used as a diagnostic tool but for its designated purpose, as a screening instrument (Stone et al., 2010). If used for epidemiological or clinical screening purposes, findings need to be viewed with caution as trauma sequelae may be missed. In clinical settings we recommend follow-up clinical interviews, to ensure accurate diagnosis (Ehnholt & Yule, 2006; Okawa, 2008). If used to evaluate interventions, the SDQ may fail to identify symptoms and detect symptom changes.

The need for qualitative studies into translated SDQs was highlighted by Derluyn and Broekaert (2007), who questioned whether refugee respondents understood the SDQ's idioms of distress. This articles' findings of structural (and hence construct) inequivalence across many countries and languages underscore the need to conduct cognitive interviews across cultures to investigate how respondents understand items, with findings incorporated into a revised, culturally sensitive SDQ. Such a project would create an opportunity for the type of cross-country research collaboration and integration proposed by the European ROAMER project (Wittchen, Knappe, & Schumann, 2014), and by the European Society for Child and Adolescent Psychiatry (ESCAP; Hebebrand et al., 2016).

ROAMER aims to improve translation of research into practice and identification and treatment of mental disorders, as exemplified in the work of Essau et al. (2012) and Ortuño-Sierra et al. (2015). Collaborative research has particular salience in the context of the current movement of child, adolescent and adult refugees from the Middle East, and other countries, and would be consistent with ESCAPs call for cross-country cooperation in screening and responding to young refugees' mental health needs (Gadeberg & Norredam, 2016; Hebebrand et al., 2016).

This review is limited by the search methods used to identify SDQs translated into refugee-relevant languages, which included restricting searches to English-language publications. Conclusions are particularly limited by the lack of studies that validated translated SDQs with children identified as refugees. By focusing on refugee-relevant translations we may have failed to identify deficits in evidence for the validity of other translated SDQs.

5 | CONCLUSION

Studies of SDQs translated into refugee-relevant languages failed to report whether challenges were experienced in establishing conceptual and linguistic equivalence with the original SDQ. Evidence was lacking for speculations regarding cultural influences on findings. Factor analysis predominantly failed to support the 5-factor model, with only two of 13 translation studies showing clear structural equivalence. This suggests translated subscales may not measure the same constructs as the UK SDQ. Translated SDQs' evidence for clinical predictive validity was equivocal, with somewhat higher (though still low) sensitivity for detecting disorders than in the United Kingdom, and somewhat lower specificity. Mean scores for translated SDQs showed considerable variation from UK means, which indicates the importance of avoiding comparisons between refugee and majority population means unless local norms are developed.

If the SDQ is used to screen for mental health problems in refugee or other trauma-affected children it should be used with caution as it does not include content that would identify the sequelae of trauma. Given the subscales' structural inequivalence, use of the SDQ TDS is recommended to obtain a broad-brush picture of the child's difficulties, with awareness that parents and adolescents may not understand the SDQ's idioms of distress, or may over- or under-state problems. Parents' and adolescents' ratings may differ from teachers' ratings as teachers may rate refugee children according to Western behavioural norms.

Following SDQ screening, a follow-up clinical interview is recommended, bearing in mind the high likelihood of false positives (or non-cases) due to the equivocal predictive equivalence of translated SDQs. Evidence supporting translated SDQs as outcome measures is extremely limited. Until qualitative studies are conducted of refugee and immigrant respondents' understanding of items, the SDQ should be used with understanding of its limitations.

DECLARATION OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

REFERENCES

- Achenbach, T. M. (1991). *Manual for Child Behavior Checklist/ 4-18 and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.
- Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, 101(2), 213–232.
- Achenbach, T. M., Becker, A., Dopfner, M., Heiervang, E., Roessner, V., Steinhausen, H.-C., & Rothenberger, A. (2008). Multicultural assessment of child and adolescent psychopathology with ASEBA and SDQ instruments: Research findings, applications, and future directions. *Journal of Child Psychology & Psychiatry*, 49(3), 251–275.
- Acquadro, C., Conway, K., Hareendran, A., & Aaronson, N. (2008). Literature review of methods to translate health-related quality of life questionnaires for use in multinational clinical trials. *Value in Health*, 11(3), 509–521 <https://doi.org/10.1111/j.1524-4733.2007.00292.x>.
- Akobeng, A. K. (2007). Understanding diagnostic tests 3: Receiver operating characteristic curves. *Acta Paediatrica*, 96(5), 644–647 <https://doi.org/10.1111/j.1651-2227.2006.00178.x>.
- Almaqrani, M. H., & Shuwail, A. Y. (2004). Validity of the self-report version of the strengths and difficulties questionnaire in Yemen. *Saudi Medical Journal*, 25(5), 592–601.
- Alyahri, A., & Goodman, R. (2006). Validation of the Arabic Strengths and Difficulties Questionnaire and the Development and Well-Being Assessment. *Eastern Mediterranean Health Journal*, 12(Suppl. 2), S138–S146.
- American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education (2014). *Standards for Educational and Psychological Testing*. Washington, DC: AERA.
- American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (fourth ed.). Washington, DC: American Psychiatric Association.
- Andary, L., Stolk, Y., & Klimidis, S. (2003). *Assessing Mental Health Across Cultures*. Brisbane: Australian Academic Press.
- Arnold, B. R., & Matus, Y. E. (2000). Test translation and cultural equivalence methodologies for use with diverse populations. In I. Cuellar, & F. A. Paniagua (Eds.), *Handbook of Multicultural Mental Health: Assessment and Treatment of Diverse Populations* (pp. 121–136). San Diego, CA: Academic Press.
- Australian Mental Health Outcomes and Classification Network. (2015). NOCC – What is it? National Outcomes and Casemix collection. <http://www.amhcn.org/background/nocc-what-it>
- Barnette, J. J. (2000). Effects of stem and Likert response option reversals on survey internal consistency: If you feel the need, there is a better alternative to using those negatively worded stems. *Educational and Psychological Measurement*, 60(3), 361–370 <https://doi.org/10.1177/00131640021970592>.
- Birleson, P., Hudson, I., Buchanan, D., & Wolff, S. (1987). Clinical evaluation of a self-rating scale for depressive disorder in childhood (depression self-rating scale). *Journal of Child Psychology and Psychiatry*, 28(1), 43–60.
- Bourdon, K. H., Goodman, R., Rae, D. S., Simpson, G., & Koretz, D. S. (2005). The strengths and difficulties questionnaire: U.S. normative data and psychometric properties. *Journal of the American Academy of Child and Adolescent Psychiatry*, 44(6), 557–564 <https://doi.org/10.1097/01.chi.0000159157.57075.c8>.
- Bowling, A. (2005). Mode of questionnaire administration can have serious effects on data quality. *Journal of Public Health*, 27(3), 281–291 <https://doi.org/10.1093/pubmed/fdi031>.
- Bronstein, I., & Montgomery, P. (2011). Psychological distress in refugee children: A systematic review. *Clinical Child and Family Psychology Review*, 14(1), 44–56 <https://doi.org/10.1007/s10567-010-0081-0>.

- Derluyn, I., & Broekaert, E. (2007). Different perspectives on emotional and behavioural problems in unaccompanied refugee children and adolescents. *Ethnicity & Health*, 12(2), 141–162 <https://doi.org/10.1080/13557850601002296>.
- Du, Y., Kou, J., & Coghill, D. (2008). The validity, reliability and normative scores of the parent, teacher and self report versions of the strengths and difficulties questionnaire in China. *Child and Adolescent Psychiatry and Mental Health*, 2(8), <https://doi.org/10.1186/1753-2000-2-8>.
- Dura-Vila, G., Klasen, H., Makatini, Z., Rahimi, Z., & Hodes, M. (2013). Mental health problems of young refugees: Duration of settlement, risk factors and community-based interventions. *Clinical Child Psychology and Psychiatry*, 18(4), 604–623 <https://doi.org/10.1177/1359104512462549>.
- Ehntholt, K. A., & Yule, W. (2006). Practitioner review: Assessment and treatment of refugee children and adolescents who have experienced war-related trauma. *Journal of Child Psychology and Psychiatry*, 47(12), 1197–1210 <https://doi.org/10.1111/j.1469-7610.2006.01638.x>.
- Elander, J., & Rutter, M. (1996). Use and development of the Rutter parents and teachers' scales. *International Journal of Methods in Psychiatric Research*, 6(2), 63–78 [https://doi.org/10.1002/\(SICI\)1234-988X\(199607\)6:2<3.3.CO;2-M](https://doi.org/10.1002/(SICI)1234-988X(199607)6:2<3.3.CO;2-M).
- Essau, C. A., Olaya, B., Anastassiou-Hadjicharalambous, X., Pauli, G., Gilvarry, C., Bray, D., ... 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>.
- Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, 6(2), 161–180 <https://doi.org/10.1037/1082-989X.6.2.161>.
- Fischer, R., & Fontaine, J. R. J. (2011). Methods for investigating structural equivalence. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Cross-cultural Research Methods in Psychology* (pp. 179–215). Cambridge: Cambridge University Press.
- Ford, T., Hutchings, J., Bywater, T., Goodman, A., & Goodman, R. (2009). Strengths and difficulties questionnaire added value scores: Evaluating effectiveness in child mental health interventions. *The British Journal of Psychiatry*, 194(6), 552–558 <https://doi.org/10.1192/bjp.bp.108.052373>.
- Fugard, A. J. B., Stapley, E., Ford, T., Law, D., Wolpert, M., & York, A. (2015). Analysing and reporting UK CAMHS outcomes: An application of funnel plots. *Child and Adolescent Mental Health*, 20(3), 155–162 <https://doi.org/10.1111/camh.12086>.
- Gadeberg, A. K., & Norredam, M. (2016). Urgent need for validated trauma and mental health screening tools for refugee children and youth. *European Child & Adolescent Psychiatry*, 25(8), 929–931 <https://doi.org/10.1007/s00787-016-0837-2>.
- Ghanizadeh, A., Mohammadi, M. R., & Yazdanshenas, A. (2006). Psychometric properties of the Farsi translation of the kiddie schedule for affective disorders and schizophrenia – Present and lifetime version. *BMC Psychiatry*, 6(10), 1–5 <https://doi.org/10.1186/1471-244X-6-10>.
- Ghanizadeh, A., Izadpanah, A., & Abdollahi, G. (2007). Scale validation of the strengths and difficulties questionnaire in Iranian children. *Iranian Journal of Psychiatry*, 2(2), 65–71.
- Goodman, R. (1997). The strengths and difficulties questionnaire: A research note. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 38(5), 581–586.
- Goodman, R. (1999). The extended version of the strengths and difficulties questionnaire as a guide to child psychiatric caseness and consequent burden. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, 40(5), 791–799.
- Goodman, R. (2001). Psychometric properties of the strengths and difficulties questionnaire. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(11), 1337–1345.
- Goodman, R., & Scott, S. (1999). Comparing the strengths and difficulties questionnaire and the child behavior checklist: Is small beautiful? *Journal of Abnormal Child Psychology*, 27(1), 17–24.
- Goodman, R., Meltzer, H., & Bailey, V. (1998). The strengths and difficulties questionnaire: A pilot study on the validity of the self-report version. *European Child & Adolescent Psychiatry*, 7(3), 125–130.
- Goodman, R., Ford, T., Richards, H., Gatward, R., & Meltzer, H. (2000). The development and well-being assessment: Description and initial validation of an integrated assessment of child and adolescent psychopathology. *Journal of Child Psychology & Psychiatry & Allied Disciplines*, 41(5), 645–655.
- Goodman, R., Ford, T., Simmons, H., Gatward, R., & Meltzer, H. (2003). Using the strengths and difficulties questionnaire (SDQ) to screen for child psychiatric disorders in a community sample. *International Review of Psychiatry*, 15(1/2), 166–172 <https://doi.org/10.1080/0954026021000046128>.
- Goodman, R., Slobodskaya, H., & Knyazev, G. (2005). Russian child mental health: A cross-sectional study of prevalence and risk factors. *European Child & Adolescent Psychiatry*, 14(1), 28–33 <https://doi.org/10.1007/s00787-005-0420-8>.
- 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, A., Heiervang, E., Fleitlich-Bilyk, B., Alyahri, A., Patel, V., Mullick, M., ... Goodman, R. (2012). Cross-national differences in questionnaires do not necessarily reflect comparable differences in disorder prevalence. *Social Psychiatry & Psychiatric Epidemiology*, 47(8), 1321–1331 <https://doi.org/10.1007/s00127-011-0440-2>.
- Grigorenko, E. L., Geiser, C., Slobodskaya, H. R., & Francis, D. J. (2010). Cross-informant symptoms from CBCL, TRF, and YSR: Trait and method variance in a normative sample of Russian youths. *Psychological Assessment*, 22(4), 893–911 <https://doi.org/10.1037/a0020703>.
- He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. Online Readings in Psychology and Culture, Unit 2. <http://scholarworks.gvsu.edu/orpc/vol2/iss2/8>
- Hebebrand, J., Anagnostopoulos, D., Eliez, S., Linse, H., Pejovic-Milovancevic, M., & Klasen, H. (2016). A first assessment of the needs of young refugees arriving in Europe: What mental health professionals need to know. *European Child & Adolescent Psychiatry*, 25(1), 1–6 <https://doi.org/10.1007/s00787-015-0807-0>.
- Henley, J., & Robinson, J. (2011). Mental health issues among refugee children and adolescents. *Clinical Psychologist*, 15(2), 51–62 <https://doi.org/10.1111/j.1742-9552.2011.00024.x>.
- Hodes, M. (2002). Three key issues for young refugees' mental health. *Transcultural Psychiatry*, 39(2), 196–213 <https://doi.org/10.1177/136346150203900206>.
- Howell, D. C. (2002). *Statistical Methods for Psychology* (fifth ed.). Pacific Grove, CA: Duxbury, Thomson Learning.
- International Test Commission (ITC). (2016). The ITC guidelines for translating and adapting tests (second edition). <http://www.psyssa.com/wp-content/uploads/2015/11/ITC-Guidelines-Translating-and-Adapting-Tests-v2-3.pdf>
- Kaplan, I. (2013). Trauma, development and the refugee experience: The value of an integrated approach to practice and research. In K. D. Gioia, & P. Whiteman (Eds.), *Children and Childhoods 3: Immigrant and Refugee Families* (pp. 1–22). Newcastle-upon-Tyne: Cambridge Scholars Publishing.
- Kashala, E., Elgen, I., Sommerfelt, K., & Tylleskar, T. (2005). Teacher ratings of mental health among school children in Kinshasa, Democratic Republic of Congo. *European Child & Adolescent Psychiatry*, 14(4), 208–215 <https://doi.org/10.1007/s00787-005-0446-y>.
- Kieling, C., Baker-Henningham, H., Belfer, M., Conti, G., Ertem, I., Omigbodun, O., ... Rahman, A. (2011). Child and adolescent mental health worldwide: Evidence for action. *The Lancet*, 378(9801), 1515–1525 [https://doi.org/10.1016/s0140-6736\(11\)60827-1](https://doi.org/10.1016/s0140-6736(11)60827-1).

- Kinzie, J. D., Manson, S. M., Vinh, D. T., Toaln, N. T., Anh, B., & Pho, T. N. (1982). Development and validation of a Vietnamese language depression rating scale. *American Journal of Psychiatry*, 139(10), 1276–1281.
- Lai, K. Y. C., Luk, E. S. L., Leung, P. W. L., Wong, A. S. Y., Law, L., & Karen, H. (2010). Validation of the Chinese version of the strengths and difficulties questionnaire in Hong Kong. *Social Psychiatry & Psychiatric Epidemiology*, 45(12), 1179–1186.
- Lalkhen, A. G., & McCluskey, A. (2008). Clinical tests: Sensitivity and specificity. *Continuing Education in Anaesthesia, Critical Care & Pain*, 8(6), 221–223 <https://doi.org/10.1093/bjaceaccp/mkn041>.
- Lawrence, D., Johnson, S., Hafekost, J., Boterhoven de Haan, K., Sawyer, M., Ainley, J., & Zubrick, S. R. (2015). The mental health of children and adolescents. Report on the second Australian child and adolescent survey of mental health and wellbeing. Canberra: Department of Health. [https://www.health.gov.au/internet/main/publishing.nsf/Content/9DA8CA21306FE6EDCA257E2700016945/\\$File/child2.pdf](https://www.health.gov.au/internet/main/publishing.nsf/Content/9DA8CA21306FE6EDCA257E2700016945/$File/child2.pdf)
- Lee, W., Jones, L., Goodman, R., & Heyman, I. (2005). Broad outcome measures may underestimate effectiveness: An instrument comparison study. *Child & Adolescent Mental Health*, 10(3), 143–144 <https://doi.org/10.1111/j.1475-3588.2005.00350.x>.
- Liu, S.-K., Chien, Y.-L., Shang, C.-Y., Lin, C.-H., Liu, Y.-C., & Gau, S. S.-F. (2013). Psychometric properties of the Chinese version of strength and difficulties questionnaire. *Comprehensive Psychiatry*, 54, 720–730 <https://doi.org/10.1016/j.comppsy.2013.01.002>.
- Lorek, A., Ehntholt, K., Nesbitt, A., Wey, E., Githinji, C., Rössor, E., & Wickramasinghe, R. (2009). The mental and physical health difficulties of children held within a British immigration detention center: A pilot study. *Child Abuse & Neglect*, 33(9), 573–585 <https://doi.org/10.1016/j.chiabu.2008.10.005>.
- Loutan, L., Bollini, P., Pampallona, S., Haan, D. B. D., & Gariazzo, F. (1999). Impact of trauma and torture on asylum-seekers. *European Journal of Public Health*, 9(2), 93–96 <https://doi.org/10.1093/eurpub/9.2.93>.
- Lukumar, P., Wijewardana, K., Hermansson, J., & Lindmark, G. (2008). Validity and reliability of Tamil version of strengths and difficulties questionnaire. *Ceylon Medical Journal*, 53(2), 48–52.
- Meade, A. W., & Lautenschlager, G. J. (2004). Same question, different answers: CFA and two IRT approaches to measurement invariance. Paper presented at the 19th Annual Conference of the Society for Industrial and Organizational Psychology, Chicago, IL. <https://pdfs.semanticscholar.org/0685/2710c39ae93bfa090fc553d06e01b9751c46.pdf>
- Mieloo, C. L., Bevaart, F., Donker, M. C. H., van Oort, F. V. A., Raat, H., & Jansen, W. (2014). 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>.
- Milfont, T. L., & Fischer, R. (2010). Testing measurement invariance across groups: Applications in cross-cultural research. *International Journal of Psychological Research*, 3(1), 111–121.
- O'Shea, B., Hodes, M., Down, G., & Bramley, J. (2000). A school-based mental health service for refugee children. *Clinical Child Psychology & Psychiatry*, 5(2), 189–201 <https://doi.org/10.1177/1359104500005002004>.
- Okawa, J. (2008). Considerations for the cross-cultural evaluation of refugees and asylum seekers. In L. A. Suzuki, & J. G. Ponterotto (Eds.), *Handbook of Multicultural Assessment: Clinical, Psychological and Educational Applications* (pp. 165–194). San Francisco, CA: John Wiley & Sons.
- Ortuño-Sierra, J., Fonseca-Pedrero, E., Aritio-Solana, R., Velasco, A. M., de Luis, E. C., Schumann, G., ... Lawrence, C. (2015). New evidence of factor structure and measurement invariance of the SDQ across five European nations. *European Child & Adolescent Psychiatry*, 24(12), 1523–1534 <https://doi.org/10.1007/s00787-015-0729-x>.
- Panther-Brick, C., Goodman, A., Tol, W., & Eggerman, M. (2011). Mental health and childhood adversities: A longitudinal study in Kabul, Afghanistan. *Journal of the American Academy of Child & Adolescent Psychiatry*, 50(4), 349–363.
- Perrin, S., Meiser-Stedman, R., & Smith, P. (2005). The children's revised impact of event scale (CRIES): Validity as a screening instrument for PTSD. *Behavioural and Cognitive Psychotherapy*, 33(4), 487–498 <https://doi.org/10.1017/S1352465805002419>.
- Plitas, A., Tucker, A., Kritikos, A., Walters, A., & Bardenhagen, F. (2009). A comparative study of the cognitive performance of Greek Australian and Greek national elderly: Implications for neuropsychological practice. *Australian Psychologist*, 44(1), 27–39 <https://doi.org/10.1080/00050060802587694>.
- Refugee Council. (2015). Multilingual information: Language profiles. http://www.refugeecouncil.org.uk/languages/language_profiles
- Refugee Council of Australia (RCA). (2016a). UNHCR global trends 2015 – How Australia compares with the world. <https://www.refugeecouncil.org.au/getfacts/statistics/unchr2015/>
- Refugee Council of Australia (RCA). (2016b). Global refugee statistics: Resettlement statistics, 2004 to 2013. <https://www.refugeecouncil.org.au/getfacts/statistics/global-refugee-statistics/>
- Richter, J., Sagatun, Å., Heyerdahl, S., Oppedal, B., & Røysamb, E. (2011). The strengths and difficulties questionnaire (SDQ) – Self-report. An analysis of its structure in a multiethnic urban adolescent sample. *Journal of Child Psychology & Psychiatry*, 52(9), 1002–1011 <https://doi.org/10.1111/j.1469-7610.2011.02372.x>.
- Rousseau, C., & Drapeau, A. (1998). Parent-child agreement on refugee children's psychiatric symptoms: A transcultural perspective. *Journal of the American Academy of Child & Adolescent Psychiatry*, 37(6), 629–636 <https://doi.org/10.1097/00004583-199806000-00013>.
- Rousseau, C., Benoit, M., Gauthier, M., Lacroix, L., Alain, N., Rojas, M. V., ... Bourassa, D. (2007). Classroom drama therapy program for immigrant and refugee adolescents: A pilot study. *Clinical Child Psychology & Psychiatry*, 12(3), 451–465 <https://doi.org/10.1177/1359104507078477>.
- Rousseau, C., Benoit, M., Lacroix, L., & Gauthier, M.-F. (2009). Evaluation of a sandplay program for preschoolers in a multiethnic neighborhood. *Journal of Child Psychology & Psychiatry*, 50(6), 743–750.
- Rousseau, C., Armand, F., Laurin-Lamothe, A., Gauthier, M.-F., & Saboundjian, R. (2012). A pilot project of school-based intervention integrating drama and language awareness. *Child & Adolescent Mental Health*, 17(3), 187–190 <https://doi.org/10.1111/j.1475-3588.2011.00629.x>.
- Rousseau, C., Beauregard, C., Daignault, K., Petrakos, H., Thombs, B. D., Steele, R., ... Hechtman, L. (2014). A cluster randomized-controlled trial of a classroom-based drama workshop program to improve mental health outcomes among immigrant and refugee youth in special classes. *PloS One*, 9(8), 1–9 <https://doi.org/10.1371/journal.pone.0104704>.
- Ruchkin, V., Koposov, R., & Schwab-Stone, M. (2007). The strength and difficulties questionnaire: Scale validation with Russian adolescents. *Journal of Clinical Psychology*, 63(9), 861–869 <https://doi.org/10.1002/jclp.20401>.
- Ruchkin, V., Koposov, R., Vermeiren, R., & Schwab-Stone, M. (2012). The strength and difficulties questionnaire: Russian validation of the teacher version and comparison of teacher and student reports. *Journal of Adolescence*, 35, 87–96 <https://doi.org/10.1016/j.adolescence.2011.06.003>.
- Samad, L., Hollis, C., Prince, M., & Goodman, R. (2005). Child and adolescent psychopathology in a developing country: Testing the validity of the strengths and difficulties questionnaire (Urdu version). *International Journal of Methods in Psychiatric Research*, 14(3), 158–166 <https://doi.org/10.1002/mpr.3>.
- Shahrivar, Z., Tehrani-Doost, M., Pakbaz, B., Rezaie, A., & Ahmadi, F. (2009). Normative data and psychometric properties of the parent and teacher versions of the strengths and difficulties questionnaire (SDQ) in an Iranian community sample. *Journal of Research in Medical Sciences*, 14(2), 69–77.
- Sireci, S. (2011). Evaluating test and survey items for bias across languages and cultures. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Cross-cultural Research Methods in Psychology* (pp. 216–243). Cambridge: Cambridge University Press.

- Stevanovic, D., Urbán, R., Atilola, O., Vostanis, P., Singh Balhara, Y. P., Avicenna, M., ... Petrov, P. (2014). 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., 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>.
- Su, L.-Y., Li, X.-R., Huang, C.-X., Luo, X.-R., & Zhang, J.-S. (2001). Norms of the Conners parent symptom questionnaire in Chinese urban children. *Chinese Journal of Clinical Psychology*, 9(4), 241–243.
- Sullivan, G. M., & Feinn, R. (2012). Using effect size – Or why the P value is not enough. *Journal of Graduate Medical Education*, 4(3), 279–282 <https://doi.org/10.4300/jgme-d-12-00156.1>.
- Swanson, J. M. (1992). *School-based Assessments and Interventions for ADD Students*. Irvine, CA: K.C. Publishing.
- Tesfaye, M., Hanlon, C., Wondimagegn, D., & Alem, A. (2009). Detecting postnatal common mental disorders in Addis Ababa, Ethiopia: Validation of the Edinburgh postnatal depression scale and Kessler scales. *Journal of Affective Disorders*, 122(1–2), 102–108 <https://doi.org/10.1016/j.jad.2009.06.020>.
- Thabet, A. A., Stretch, D., & Vostanis, P. (2000). Child mental health problems in Arab children: Application of the strengths and difficulties questionnaire. *International Journal of Social Psychiatry*, 46(4), 266–280 <https://doi.org/10.1177/002076400004600404>.
- United Nations High Commissioner for Refugees (UNHCR). (2015). Global trends: Forced displacement in 2015. <http://www.unhcr.org/global-trends-2015.html>
- United Nations High Commissioner for Refugees (UNHCR). (2016). Global resettlement statistical report 2014. <http://www.unhcr.org/52693bd09.pdf>
- Van de Vijver, F. J. R., & Leung, K. (2011). Equivalence and bias: A review of concepts, models, and data analytic procedures. In D. Matsumoto, & F. J. R. van de Vijver (Eds.), *Cross-cultural Research Methods in Psychology* (pp. 17–45). Cambridge: Cambridge University Press.
- Van de Vijver, F. J. R., & Poortinga, Y. H. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-cultural Assessment* (pp. 39–63). Mahwah, NJ: Lawrence Erlbaum.
- Vostanis, P. (2006). Strengths and difficulties questionnaire: Research and clinical applications. *Current Opinion in Psychiatry*, 19(4), 367–372.
- Whitehouse, W. G., Orne, E. C., & Dinges, D. F. (2002). Demand characteristics: Toward an understanding of their meaning and application in clinical practice. *Prevention & Treatment*, 5(1), 1–6 <https://doi.org/10.1037/1522-3736.5.1.534i>.
- Wiegiersma, P. A., Stellinga-Boelen, A. A. M., & Reijneveld, S. A. (2011). Psychosocial problems in asylum seekers' children: The parent, child, and teacher perspective using the strengths and difficulties questionnaire. *Journal of Nervous and Mental Disease*, 199(2), 85–90 <https://doi.org/10.1097/NMD.0b013e31820446d2>.
- Williamson, A., McElduff, P., Dadds, M., D'Este, C., Redman, S., Raphael, B., ... Eades, S. (2014). The construct validity of the strengths and difficulties questionnaire for Aboriginal children living in urban new South Wales, Australia. *Australian Psychologist*, 49(3), 163–170 <https://doi.org/10.1111/ap.12045>.
- Wittchen, H.-U., Knappe, S., & Schumann, G. (2014). The psychological perspective on mental health and mental disorder research: Introduction to the ROAMER work package 5 consensus document. *International Journal of Methods in Psychiatric Research*, 23(Suppl. 1), 15–27 <https://doi.org/10.1002/mpr.1408>.
- Woerner, W., Nuanmanee, S., Becker, A., Wongpiromsarn, Y., & Mongkol, A. (2011). Normative data and psychometric properties of the Thai version of the strengths and difficulties questionnaire (SDQ). *Journal of Mental Health of Thailand*, 19(1), 42–58.
- World Health Organization (WHO) (1992). *The ICD-10 Classification of Mental and Behavioural Disorders*. Geneva: WHO.
- World Health Organization (WHO). (2015). Management of Substance Abuse: Process of translation and adaptation of instruments. http://www.who.int/substance_abuse/research_tools/translation/en/
- Yao, S., Zhang, C., Zhu, X., Jing, X., McWhinnie, C. M., & Abela, J. R. Z. (2009). Measuring adolescent psychopathology: Psychometric properties of the self-report strengths and difficulties questionnaire in a sample of Chinese adolescents. *Journal of Adolescent Health*, 45(1), 55–62 <https://doi.org/10.1016/j.jadohealth.2008.11.006>.
- youthinmind. (2012a). Selected SDQ publications. <http://www.sdqinfo.com/py/sdqinfo/f0.py>
- youthinmind. (2012b). What is the SDQ? <http://www.sdqinfo.com/a0.html>
- youthinmind. (2013). SDQ normative data. <http://www.sdqinfo.com/g0.html>
- youthinmind. (2015). Downloadable SDQs and related items. <http://www.sdqinfo.com/py/sdqinfo/b0.py>
- Zwirs, B., Burger, H., Schulpen, T., Vermulst, A. A., HiraSing, R. A., & Buitelaar, J. (2011). Teacher ratings of children's behavior problems and functional impairment across gender and ethnicity: Construct equivalence of the strengths and difficulties questionnaire. *Journal of Cross-Cultural Psychology*, 42(3), 466–481 <https://doi.org/10.1177/0022022110362752>.

How to cite this article: Stolk Y, Kaplan I, Szwarc J. Review of the strengths and difficulties questionnaire translated into languages spoken by children and adolescents of refugee background. *Int J Methods Psychiatr Res*. 2017;26:e1568. <https://doi.org/10.1002/mpr.1568>