Child: care, health and development

Review

doi:10.1111/cch.12397

The psychometric properties of the Ages & Stages Questionnaires for ages 2-2.5: a systematic review

T. Velikonja,* J. Edbrooke-Childs,* A. Calderon,* M. Sleed,* A. Brown† and J. Deighton*

*Evidence Based Practice Unit, University College London and Anna Freud Centre, London, UK, and
†School of Psychology, University of Kent, Kent, UK

Accepted for publication 31 July 2016

Abstract

Background Early identification of children with potential development delay is essential to ensure access to care. The Ages & Stages Questionnaires (ASQ) is used as population outcome indicators in England as part of the 2.5-year review.

Method The aim of this article was to systematically review the worldwide evidence for the psychometric properties of the ASQ third edition (ASQ-3™) and the Ages & Stages Questionnaires®: Social-Emotional (ASQ:SE). Eight electronic databases and grey literature were searched for original research studies available in English language, which reported reliability, validity or responsiveness of the ASQ-3™ or ASQ:SE for children aged between 2 and 2.5 years. Twenty studies were included. Eligible studies used either the ASQ-3™ or the ASQ:SE and reported at least one measurement property of the ASQ-3™ and/or ASQ:SE. Data were extracted from all papers identified for final inclusion, drawing on Cochrane guidelines.

Results Using 'positive', 'intermediate' and 'negative' criteria for evaluating psychometric properties, results showed 'positive' reliability values in 11/18 instances reported, 'positive' sensitivity values in 13/18 instances reported and 'positive' specificity values in 19/19 instances reported.

Conclusions Variations in age or language versions used, quality of psychometric properties and quality of papers resulted in heterogeneous evidence. It is important to consider differences in cultural and contextual factors when measuring child development using these indicators. Further research is very likely to have an important impact on the interpretation of the ASQ-3™ and ASQ:SE psychometric evidence.

Keywords
ASQ, developmental delay, population outcome indicator, psychometric

Correspondence: Julian Edbrooke-Childs, UCL and the Anna Freud Centre, 12 Maresfield Gardens, London NW3 5SD, UK
E-mail: ebpu@annafreud.org

Introduction

Early identification of developmental problems and disabilities is essential to increase access to evaluation and intervention (Briggs-Gowan & Carter 2008). Evidence suggests that without appropriate support, early difficulties are resistant to change and are even likely to intensify over time (Feil *et al.* 1998). The monitoring of child development has a pivotal role in paediatric care (Heo & Squires 2012; Sheldrick & Perrin 2013), as early identification and intervention may influence the course of otherwise persistent difficulties (Brugman *et al.* 2001; Briggs-Gowan & Carter 2008).

In 2012/2013, with a view to developing a public health outcome measure for children aged 2–2.5, the Department of Health (DH) in England commissioned a review of various existing measures of early development. The measure would be used to monitor child development across England, with the following aims: (i) observe changes in population health over time; (ii) track children's outcomes as they grow up; (iii) evaluate the effectiveness of services for 0–2 year olds, supporting planning; and (iv) assist health visitors with identification (and intervention) of children's early developmental problems (DH 2014).

The two-phased review (Bedford *et al.* 2013; Kendall *et al.* 2014) identified the ASQ-3^{TM} as a measure of child development best fitting the two main DH prerequisites: the inclusion of all aspects of child development (physical, social, emotional, cognitive, and speech and language) and the ability to be applied as a population outcome measure.

Based on these findings, our objective was to examine studies published worldwide relating to the validity and reliability of the ASQ-3^{TM} and ASQ:SE, and seek to draw conclusions for the English context.

Background of the ASQ

The ASQ were developed in the 1980s by Jane Squires and Diane Bricker at the University of Oregon. After years of refinements, the questionnaires were published in 1995 as 'Ages & Stages Questionnaires® (ASQ): A Parent-Completed, Child-Monitoring System'. The third edition (i.e. ASQ-3^{TM}) was published in 2009. The ASQ-3^{TM} was designed to identify potential developmental delay in children aged between one month and 5.5 years in five domains (communication, gross motor, fine motor, problem-solving and personal–social). It has been used for research and in clinical contexts across disciplines; e.g. medical settings (Pinto-Martin *et al.* 2005; Council on Children with Disabilities *et al.* 2006) and early intervention services (Baggett *et al.* 2007; Flamant *et al.* 2011). As well as its use in North America, it has been translated and used around the world; e.g. Europe (Kerstjens *et al.* 2009; Sarmiento Campos *et al.* 2011; Troude *et al.* 2011; Østergaard *et al.* 2012; Sidor *et al.* 2013; Lopes *et al.* 2014), Asia (Bian *et al.* 2010; Saihong 2010; Bian *et al.* 2012; Heo & Squires 2012; Juneja *et al.* 2012), South America (Filgueiras *et al.* 2013; Schonhaut *et al.* 2013) and Australia (D'Aprano *et al.* 2014).

The Ages & Stages Questionnaires®: Social–Emotional (ASQ:SE): A Parent-Completed Child Monitoring System for Social–Emotional Behaviours was developed to be used alone or in conjunction with the ASQ-3^{TM} (or other developmental measures), and it focuses on infants' and young children's social and emotional development.

There is no definitive test of developmental progress in early childhood as there is wide variation in what can be considered typical at any one age, and the factors associated with developmental difficulties may be complex in both aetiology and prognosis. However, while there is no objective 'gold standard', psychometric instruments do exist that have established themselves as trusted measures of various types of delay. Comparing the ASQ to the most well-established measures is important for understanding its comparative value. In terms of cognitive-motor development, the Bayley Scales of Infant Development (Bayley 1993), completed by professionals, can be used with children of up to 3.5 years and may be considered the closest comparator; however, some evidence has questioned its sensitivity and predictive validity (Moore *et al.* 2012; Luttikhuizen dos Santos *et al.* 2013; Spittle *et al.* 2013). For socio-emotional development, the Child Behavior Checklist (Achenbach 1992) may be considered the closest comparator; it is completed by parents or teachers for children under 11 years and has evidence of high sensitivity and predictive validity (Verhulst *et al.* 1994; Mick *et al.* 2003).

This research aimed to systematically review international evidence regarding the psychometric properties of the ASQ (ASQ-3^{TM} and ASQ:SE) for ages 2–2.5 (24-, 27- and 30-month versions of the questionnaires). This is to inform the use of the ASQ as population outcome indicators in England at 2.5 years, the age at which children are reviewed using these measures (Bedford *et al.* 2013; Kendall *et al.* 2014).

Methods

The systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) (Moher *et al.* 2009). Research questions, objectives, methods of analysis and inclusion criteria were specified in advance and documented in a protocol.

Inclusion criteria

The inclusion and exclusion criteria and search strategy were agreed with the advisory team members, which included 11 experts in child development and psychological measurement. All included studies were original research papers, written in English, published between 1995[1] (the year questionnaires were first published) and 15 December 2014. All language versions of the ASQ-3^{TM} and ASQ:SE were included because as

[1] 1995 was selected to provide DH with more detailed information about studies using any versions of the ASQ as population outcome measures, but goes beyond the scope of this paper. This paper only focuses on papers published in/after 2009, the year when the ASQ-3 was first published.

the original instruments are from the United States it was important to explore how the measure has been translated or adapted to other contexts, and how the psychometric properties have been affected when doing so. For this paper, only the latest edition of the ASQ-3^{TM} was considered, as comparison across all versions of the measure would not be feasible because of revisions (e.g. new open-ended questions, new standardization, revised cut-off points, new 'monitoring zone').

Studies were eligible if they used either the ASQ-3^{TM} (24-, 27- or 30-month version – chosen to correspond to the age at which children are reviewed in England) or the ASQ:SE (24- or 30-month versions), reported one or more measurement property of the ASQ-3^{TM} and/or ASQ:SE, and included information on the study design and data analysis procedure used to allow completion of the COnsensus-based Standards for the selection of health Measurement Instruments (COSMIN) 4-point checklist (COSMIN group), criteria against which the quality of reporting studies of psychometric properties may be assessed.

Search strategy

A literature search was conducted in eight databases: PsycINFO, PubMed, Web of Science, EMBASE, Health and Psychosocial Instruments (HaPI), ERIC, The Cochrane Library and CINAHL Plus. This selection was based on: the COSMIN guidelines, the broader topics that the review covered and existing systematic reviews of similar instruments (McCullough & Parkes 2008; Eeles *et al.* 2013; Field & Livingstone 2013).

Grey literature outside commercial or academic publishing was also included, using these databases: Index to Theses, Dissertation and Theses, PsycEXTRA and OpenSIGLE. Also, the Ages and Stages website (*agesandstages.com*) was reviewed for reports. Individuals known to have relevant expertise were contacted to gather knowledge of any ongoing, as-yet-unpublished research. '*ASQ around the World*' Symposium (*San Francisco*, *September 2014*) presentations were included and contributors contacted to gather some of the most up-to-date research data on psychometric properties and utility of the ASQ-3^{TM} and ASQ:SE. Additional searches for further evidence were completed (e.g. citation tracking of identified papers, Google Scholar search, searching relevant journals). Any irretrievable papers were sought by direct email contact of the first two authors of each manuscript. Search terms can be found in the appendix.

Data screening and extraction

Two reviewers carried out filtering in parallel. For the first screening, the second reviewer completed approximately 10% of all included papers ($n = 620$). The inter-rater reliability was 0.80 (Kappa value), which signifies a very good level of agreement (Landis & Koch 1977). For the full-text screening, both reviewers screened all papers, with the inter-rater reliability 0.83. Any reviewer discrepancies were discussed, in all cases resulting in exclusion of these papers as not relevant.

Data were extracted from all included papers, drawing on Cochrane guidelines (Higgins & Green 2011). The initial framework against which the nature and quality of the evidence provided was assessed drew on Terwee *et al.*'s (2007) criteria for appraising measurement properties of questionnaires, and was modified as necessary to meet this study's aims. All values of psychometric properties were transformed to 'positive' (+), 'intermediate' (+/−) or 'negative' (−) following the criteria[2] (*e.g. Cronbach's alpha of above 0.70 is considered a 'positive' value*).

Quality assessment

In addition to the Terwee *et al.* criteria, the evidence quality was assessed using an adapted version of the COSMIN checklist. The original checklist contains nine domains (e.g. internal consistency, reliability), with 5–18 items per domain. The only domain adapted was the 'cross-cultural validity' (see Schellingerhout *et al.* 2011 and Appendix 1). For each item in the checklist, specific criteria were developed for 'excellent', 'good', 'fair' and 'poor' quality. An overall score for the study's methodological quality of any of the measurement properties is obtained by taking the lowest score for any of a domain's individual items. For example, if for a reliability study one item in the 'reliability' domain is scored poor, the overall methodological quality of that reliability study is rated as poor. The quality of the translation of the ASQ-3^{TM} or ASQ:SE was assessed (where applicable) using the adapted 'cross-cultural validity' items, consistent with previous reviews by COSMIN developers (Schellingerhout *et al.* 2011). Most of the grey literature was not assessed with COSMIN because the

[2] For Internal consistency (Cronbach's alpha): (−), <0.60; (+/−), 0.60–0.70; (+),>0.70; test–retest reliability (ICC): (−), <0.60; (+/−), 0.60–0.80; (+), >0.80; inter-rater reliability (ICC): (−), <0.50; (+/−), 0.50–0.70; (+), >0.70; sensitivity/specificity: (−), <0.50; (+/−), 0.50–0.70; (+), >0.70 (Terwee *et al.* 2007)

limited information about the study design and other methodological considerations was available.

Both reviewers completed the quality assessment and evaluation of the psychometric properties; any discrepancies were identified and discussed.

Results

The academic searches resulted in 6208 hits (see Fig. 1). After excluding duplicates, 4476 were identified for initial screening. Through title/abstract screening, 342 potentially relevant articles were identified. After screening, 13 studies were included.

The grey literature search returned 822 hits (see Fig. 2). After review of abstracts/titles, 29 articles were identified for full-text screening. A total of five articles were included (two technical reports and three symposium abstracts/presentations).

Study characteristics

Tables 1 and 2 summarize the study characteristics. Total sample sizes varied extensively, from 60 (Saihong 2014) to 45 640 (Filgueiras 2014), but most (72%, $n = 13$) ranged between approximately 100 and 3000 participants. Studies comprised convenience samples (Pomes 2013; Filgueiras 2014; Saihong 2014; Veldhuizen *et al.* 2014), at-risk groups (San Antonio *et al.* 2014), non-representative samples (Ivey-Soto

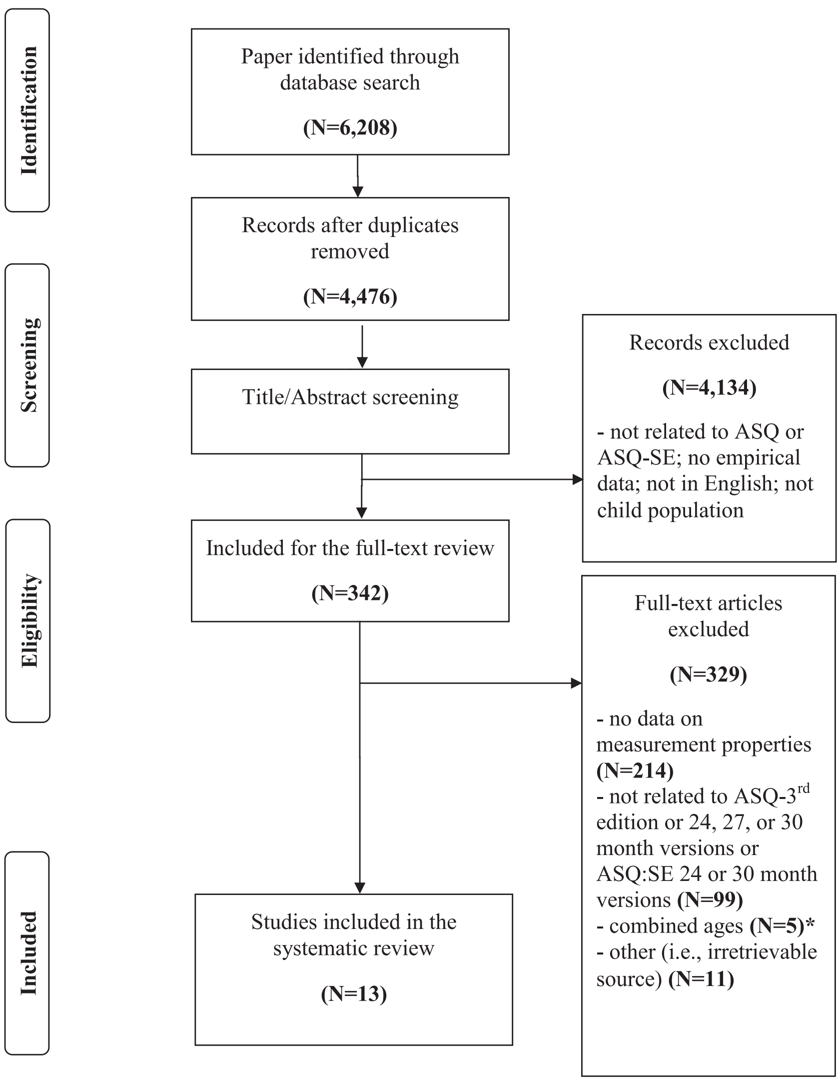

Figure 1. Main databases—Flowchart of studies included in the literature review; adapted (Eeles *et al.* 2013). *Note*: *authors contacted to obtain data for the individual ASQ-3™ and ASQ:SE age versions; if available the paper was included in the systematic review (one paper).

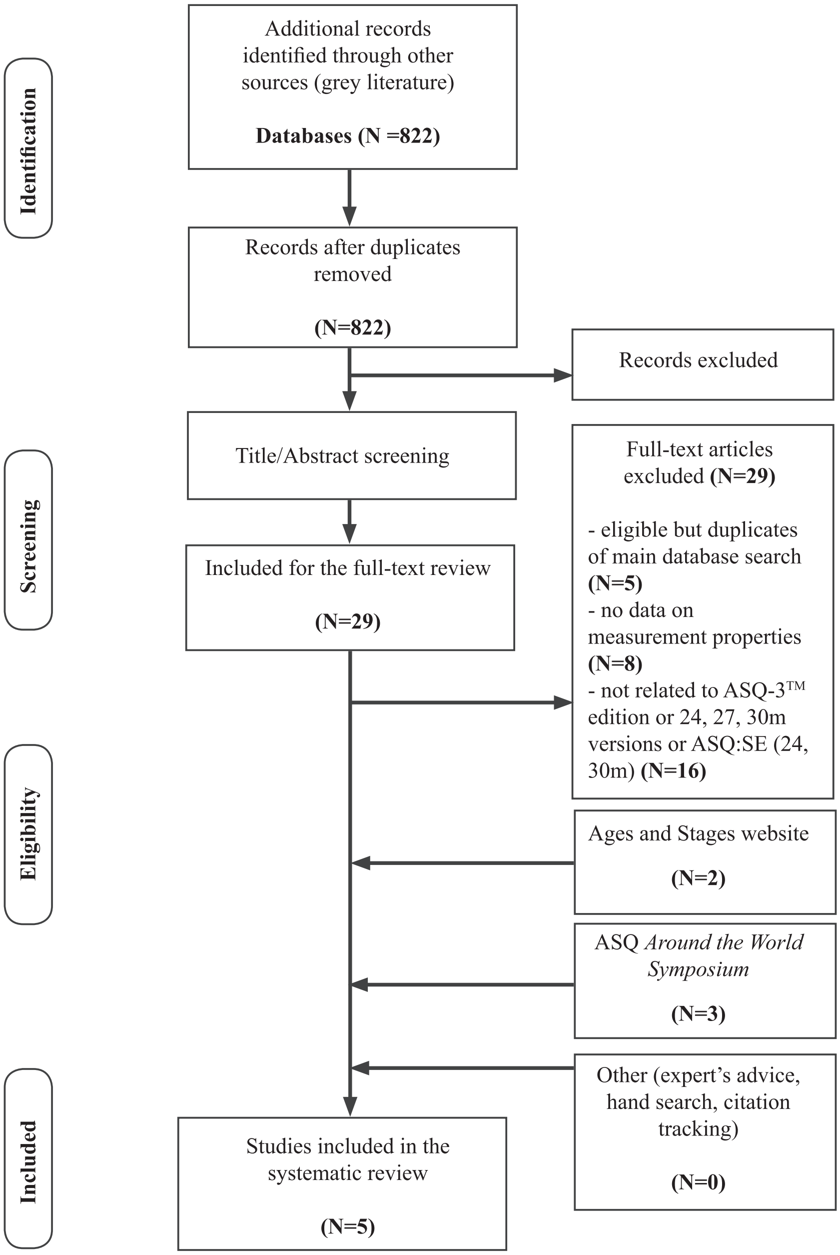

Figure 2. Grey literature—Flowchart of studies included in the literature review; adapted (Moher *et al.* 2009).

2008; Kucuker *et al.* 2011), stratified random samples (Heo 1999; Squires *et al.* 2001a; Squires *et al.* 2009a; Bian *et al.* 2012; Heo & Squires 2012; Filgueiras 2014) and representative samples (de Wolff *et al.* 2013; Filgueiras *et al.* 2013; Kvestad *et al.* 2013; Schonhaut *et al.* 2013; Lopes *et al.* 2014).

Thirty-nine percent ($n = 7$) of studies were based in North America and 51% elsewhere ($n = 11$, e.g. China, Brazil and South Korea). Sixty-one percent ($n = 11$) reported on the psychometric properties of the ASQ-3^{TM}, and 39% ($n = 7$) reported the psychometric properties of the ASQ:SE.

Reliability and validity of ASQ-3^{TM}

Table 3 summarizes the evidence found for the psychometric properties of the ASQ-3^{TM} and ASQ:SE which – when values for the total score were not available – included median values

Table 1. Characteristics of included studies

First author, year	Country	Study design	N	Questionnaires version	Age version	Aim(s) of the study
San Antonio *et al.* (2014)	USA	Convenience sample, at-risk groups	131	ASQ, English & Spanish	30 (*n* = 65)	To examine the reproducibility of the ASQ-3™ under standardized versus non-standardized conditions
Veldhuizen *et al.* (2014)	Canada	Convenience sample from community organizations	587	ASQ, English	24 (*n* = 64) 30 (*n* = 49)	To evaluate the agreement between the ASQ-3™ and the Bayley-III[b]
Heo (1999)	USA	Stratified sample	447	SE, English	24 (*n* = 237)	To examine the validity and reliability of the ASQ:SE
Ivey-Soto (2008)[†]	USA	Non-representative sample (Early Head Start programs attendees)	50	SE, English	24 (*n* = 16) 30 (*n* = 13)	To assess the utility and usability of the SEAM[a] To evaluate the relationship between the SEAM and ASQ:SE
Bian *et al.* (2012)	China	Cross-sectional study (stratified sampling)	8,472	ASQ, Chinese	24, 30	To evaluate the reliability and validity of the Chinese translation of the ASQ-3™ to identify developmental delays in preschool children
Filgueiras *et al.* (2013)	Brazil	All public day centres in Rio de Janeiro	45,640	ASQ, Portuguese (Brazilian)	24 (*n* = 1454) 27	To translate and adapt the ASQ-3™ for use in Brazilian public child day care centres
					(*n* = 2222) 30 (*n* = *2814*)	To explore the psychometric properties of the Brazilian ASQ-3™
Kvestad *et al.* (2013)	India	Sampled part of a randomized double blind placebo control study (last 440 enrolled children – out of 1000 in total – included in this study)	422	ASQ, Indian	24 (*n* = *39*) 27 (*n* = *47*) 30 (*n* = *37*)	To assess the feasibility of the ASQ-3™ 'home procedure' in epidemiological studies
Pomes (2013)[†]	USA	Convenience sample	798	ASQ, Spanish (Chilean)	30 (*n* = 177 Spanish and *n* = 127 English)	To evaluate the cultural appropriateness of the Spanish translation of the ASQ:SE To test the functional invariance of the ASQ:SE across the Spanish and English version of the ASQ-3™
Schonhaut *et al.* (2013)	Chile	Representative sample, preterm and term children, recruited from ambulatory well-child clinic	306	ASQ, Spanish (Chilean)	30 (*n* = 96)	To assess the concurrent validity of the ASQ-3™ Chile –compared with The Bayley Scales of Infant and Toddler Development (Bayley-III)
de Wolff *et al.* (2013)	Netherlands	Cross-sectional representative sample	2106	SE, Dutch	24 (*n* = 840)	To compare the psychometric properties of the three questionnaires to detect psychosocial problems in toddlers
Heo and Squires (2012)	Korea	Stratified sample (Korean census data)	2562	SE, Korean	24 (*n* = 293) 30 (*n* = 206)	To investigate the appropriateness of the translation of the Korean ASQ:SE and its validity and reliability

(continues)

Table 1. (Continued)

First author, year	Country	Study design	N	Questionnaires version	Age version	Aim(s) of the study
Kucuker *et al.* (2011)	Turkey	Non-representative sample (Preschools, child psychiatry and paediatric clinics and hospitals, special education schools, community clinics in Ankara and Denizli)	608	SE, Turkish	24 ($n = 30$) 30 ($n = 41$)	To evaluate the validity and reliability of the instrument to screen children who are (or are not) at risk of social-emotional problems

Note. 24 = ASQ-3™ – 24-month questionnaires, 27 = ASQ-3™ – 27-month questionnaires, 30 = ASQ-3™ – 30-month questionnaires.
ASQ = ASQ-3™; SE = ASQ:SE.
[a] The Social Emotional Assessment Measure (SEAM).
[b] The Bayley Scales of Infant Development, third edition (BSID).
† = Study not included in the tables of results (3 and 4) as their analytic strategy differed from other studies and, therefore, was not amenable to presentation in the same format. Sample size (reported as *n*) for different age versions of ASQ/ASQ:SE was not available for some studies.

Table 2. Characteristics of the grey literature included

First author, year	Country	Study design	*N*	Questionnaires version	Age version	Aim(s) of the study
Technical reports						
Squires *et al.* (2001a, 2009a)	USA	Stratified sample of US population	18 572	ASQ, English	24, 27, 30	To assess the validity and reliability of the ASQ-3™
Squires *et al.* (2001b)	USA	Stratified sample of US population	3014	SE, English	24, 30	To assess the validity and reliability of the ASQ:SE
ASQ Symposium abstracts/presentations						
Filgueiras (2014)	Brazil	Convenience sample	150	ASQ, Portuguese (Brazilian)	24 ($n = 50$) 27 ($n = 50$) 30 ($n = 50$)	To assess the test–retest reliability of the Brazilian ASQ-3™ using teachers' responses
Lopes *et al.* (2014)	Portugal	General population sample	1908	ASQ, Portuguese	24 ($n = 111$) 27 ($n = 109$) 30 ($n = 112$)	Evaluate the psychometric properties of the ASQ-3™ Portuguese
Saihong (2014)	Thailand	Convenience sample	60	ASQ, Thai	24 ($n = 30$) 30 ($n = 30$)	To assess the concurrent validity of the Thai ASQ-3™ compared with the Denver Development Screening II (DDST-II)

Note. 24 = ASQ-3™ – 24-month questionnaires, 27 = ASQ-3™ – 27-month questionnaires, 30 = ASQ-3™ – 30-month questionnaires.
ASQ = ASQ-3™; SE = ASQ:SE. Sample size (reported as *n*) for different age versions of ASQ/ASQ:SE was not available for some studies.

Table 3. Peer-reviewed evidence of the reliability and validity of the ASQ-3^{TM} and ASQ:SE and their translated/adapted versions

			Reliability indicators									Validity indicators					
			Internal consistency			Test–retest			Inter-rater			Sensitivity			Specificity		
First author, year	Country	Type	24	27	30	24	27	30	24	27	30	24	27	30	24	27	30
San Antonio *et al.* (2014)	USA	ASQ	?	?	?	?	?	0.84**	?	?	?	?	?	?	?	?	?
Veldhuizen *et al.* (2014)	Canada	ASQ	?	?	?	?	?	?	?	?	?	0.83** †	?	0.33**†	0.84**†	?	0.87**†
Heo (1999)	USA	SE	0.71*	n/a	?	1*	n/a	?	?	?	?	?	n/a	?	0.95***†	n/a	?
Bian *et al.* (2012)	China	ASQ:T	?	?	?	?	?	?	?	?	?	0.50****† 0.80****	?	1****†	0.89****† 0.84***	?	0.85****†
Filgueiras *et al.* (2013)	Brazil	ASQ:T	0.65****	0.63****	0.70****	?	?	?	?	?	?	?	?	?	?	?	?
Kvestad *et al.* (2013)	India	ASQ:T	0.82*	0.84*	0.84*	?	?	?	?	?	?	?	?	?	?	?	?
Schonhaut *et al.* (2013)	Chile	ASQ:T	?	?	?	?	?	?	?	?	?	?	?	0.82***†	?	?	0.84***†
de Wolff *et al.* (2013)	Netherlands	SE:T	0.62*	n/a	?	?	n/a	?	?	?	?	0.66***†	n/a	?	0.91***†	n/a	?
Heo and Squires (2012)	Korea	SE:T	?	n/a	?	?	n/a	?	?	?	?	1***†	n/a	0.25***†	0.87***†	n/a	0.80***†
Kucuker *et al.* (2011)	Turkey	SE: T	0.76*	n/a	0.85*	0.67*	n/a	0.80*	?	?	?	0.90**	n/a	0.78**	0.95**	n/a	0.74**

Internal consistency (Cronbach's alpha): (−), $<$0.60; (+/−), 0.60–0.70; (+),$>$0.70; test–retest reliability (ICC): (−), $<$0.60; (+/−), 0.60−0.80; (+), $>$0.80; inter-rater reliability (ICC): (–), $<$0.50; (+/−), 0.50−0.70; (+), $>$0.70; sensitivity/specificity: (−), $<$0.50; (+/−), 0.50–0.70; (+), $>$0.70 (Terwee *et al.* 2007).
COSMIN: no asterisk = indeterminate (grey literature).
*1 = poor.
**2 = fair.
***3 = good.
****4 = excellent.
Version: 24 = ASQ 24-month questionnaire; 27 = ASQ 27-month questionnaire; 30 = ASQ 30-month questionnaire.
Type: ASQ = ASQ-3^{TM}; ASQ:T = ASQ-3^{TM}, translated; SE = ASQ:SE; SE:T = ASQ:SE, translated.
† Reference to a well-established comparable measure (CBCL or Bayley Scales).
? No information provided, n/a = does not apply.

of subscales' internal consistency reliability (Cronbach's alpha), test–retest reliability, inter-rater reliability, sensitivity and specificity.

The three ASQ-3^{TM} age versions were found to have 'positive' values for internal consistency reliability (Cronbach's alpha >0.70) based on the medians of the five ASQ-3^{TM} subscales (i.e. communication, gross motor, fine motor, problem-solving and personal–social) (Squires *et al.* 2009a). However, there was variation within specific subdomains and lower Cronbach's alpha values were found for fine-motor skills at 24 months (0.51), problem-solving at 24 months (0.53) and personal–social at 27 months (0.58) (Squires *et al.* 2009a). This was the only study (obtained from grey literature) which reported the internal consistency of the subscales using the original ASQ-3^{TM} (see Table 4).

The internal consistency reliability of the translated/adapted versions of the ASQ-3^{TM} was generally lower but consistent across the different age versions of the measure: Cronbach's alpha ranged between 0.46 (Lopes *et al.* 2014) and 0.82 (Kucuker *et al.* 2011) for the 24-month version, between 0.57 (Lopes *et al.* 2014) and 0.84 (Kucuker *et al.* 2011) for the 27-month version and between 0.52 (Lopes *et al.* 2014) and 0.84 (Kucuker *et al.* 2011) for the 30-month version. The quality of the studies varied from 'poor' (Kucuker *et al.* 2011) to 'excellent' (Filgueiras *et al.* 2013).

One study (Heo 1999) reported test–retest reliability for the ASQ-3^{TM}, and it only provided information for the 30-month version. San Antonio et al.'s results (San Antonio *et al.* 2014) showed 'positive' values for test–retest reliability across all five ASQ domains with a median Intraclass Correlation Coefficient (ICC) value of 0.84 and a 'fair' quality (COSMIN).

The three age versions for the adapted/translated ASQ-3^{TM} showed 'positive' values in two unpublished studies: Spearman's mean correlation of 0.72 (Filgueiras 2014) and Pearson product–moment correlation coefficient of 0.90 (Lopes *et al.* 2014). The time lag between measurements was two weeks in both studies.

There were no studies that assessed the inter-rater reliability of the ASQ-3^{TM} and only one unpublished study (Lopes *et al.* 2014) that examined the inter-rater reliability of the translated/adapted ASQ-3^{TM}. Lopes *et al.*'s (2014) findings showed 'excellent' (COSMIN) inter-observer values, which were consistent across all three age versions (Pearson product–moment correlation coefficient, $M_{24m} = 0.94$; $M_{27m} = 0.84$; $M_{30m} = 0.91$).

Sensitivity values for the ASQ-3^{TM} were in general 'positive'. For the 24-month version, values ranged from 0.78 (Sheldrick & Perrin 2013) to 0.91 (Squires *et al.* 2009a). Also, when compared with the established reference standard (the Bayley Scales of Infant Development – Third Edition [BSID-III] (Squires *et al.* 2009a)), 'positive' values were observed (0.83) (Veldhuizen *et al.* 2014). The studies' quality, when assessed, was 'fair', but it was not possible to assess in one study (Squires *et al.* 2009a). For the 27-month version, only one study reported sensitivity (Squires *et al.* 2009a), with a 'positive' value of 0.78. For the 30-month version, Squires *et al.* (2009a) reported a value of 0.87. However, when compared with the BSID the value dropped to 0.33 (Veldhuizen *et al.* 2014). Again, the quality of the studies was 'fair' in one study (Veldhuizen *et al.* 2014) and not possible to assess in one study (Squires *et al.* 2009a).

Sensitivity values for the adapted/translated ASQ-3^{TM} were less consistent. For the 24-month version, values ranged from 0.80 (Bian *et al.* 2012; using Denver Developmental Screening Test-Second Edition as comparator, Frankenburg *et al.* 1990) to 0.88 (Saihong 2014), but dropped to 0.50 (Bian *et al.* 2012) when compared with the established comparator (The Bayley Scales of Infant Development – Second Edition [BSID-II] (Moore *et al.* 2012)). The quality of the Bian *et al.* (2012) study studies was 'excellent', but the quality of Saihong (2014) was not possible to assess. For the 30-month version, the value was 0.54 (Saihong 2014) and increased to 0.82 (Schonhaut *et al.* 2013) when compared with the BSID-III and to 1.0 (Bian *et al.* 2012) when compared with the BSID-II. The quality of studies ranged from 'good' (Schonhaut *et al.* 2013) to 'excellent' (Bian *et al.* 2012). There was no evidence available for the 27-month version.

Specificity values for the ASQ-3^{TM} were 'positive'. For the 24-month version, a value of 0.72 was reported (Squires *et al.* 2009a) and remained 'positive' when compared with BSID-III (0.84) (Veldhuizen *et al.* 2014). For the 27-month version, the value was also 'positive' (0.86) (Squires *et al.* 2001b). For the 30-month version, a value of 0.93 (Squires *et al.* 2009a) was reported and remained 'positive' when compared with BSID-III (0.87) (Veldhuizen *et al.* 2014). When it was possible to assess (Veldhuizen *et al.* 2014), the quality of studies was 'fair' (COSMIN).

Similar 'positive' and consistent values were found for the adapted/translated ASQ-3^{TM} versions. For the 24-month version, values ranged from 0.71 (Saihong 2014) to 0.84 (Bian *et al.* 2012) and increased to 0.89 (Bian *et al.* 2012) when compared with the BSID-II. For the 30-month version, the value was 0.91 (Saihong 2014); this slightly decreased to 0.84 (Schonhaut *et al.* 2013) when compared with the BSID-III and to 0.85 (Bian *et al.* 2012) when compared with the BSID-II. There were no available data for the 27-month version. The quality of the studies ranged from 'good' to 'excellent'; it was not possible to assess quality in one study (Saihong 2014) .

Table 4. Grey evidence of the reliability and validity of the ASQ-3™ and ASQ:SE and their translated/adapted versions

			Reliability indicators									Validity indicators					
			Internal consistency			Test–retest			Inter-rater			Sensitivity			Specificity		
First author, year	Country	Type	24	27	30	24	27	30	24	27	30	24	27	30	24	27	30
Squires *et al.* (2009a)	USA	ASQ	0.77	0.78	0.78	?	?	?	?	?	?	0.91	0.78	0.87	0.72	0.86	0.93
Squires *et al.* (2001b)	USA	SE	0.80*	n/a	0.88*	?	n/a	?	?	?	?	0.71**†	n/a	0.80**†	0.93**†	n/a	0.89**†
Filgueiras (2014)	Brazil	ASQ:T	?	?	?	0.73	0.73	0.71	?	?	?	?	?	?	?	?	?
Lopes *et al.* (2014)	Portugal	ASQ:T	0.46	0.57	0.52	0.94	0.89	0.95	0.94	0.89	0.95	?	?	?	?	?	?
Saihong (2014)	Thailand	ASQ:T	?	?	?	?	?	?	?	?	?	0.88	?	0.54	0.71	?	0.91

Internal consistency (Cronbach's alpha): (−), <0.60; (+/−), 0.60–0.70; (+),>0.70; test–retest reliability (ICC): (−), <0.60; (+/−), 0.60–0.80; (+), >0.80; inter-rater reliability (ICC): (−), <0.50; (+/−), 0.50–0.70; (+), >0.70; sensitivity/specificity: (−), <0.50; (+/−), 0.50–0.70; (+), >0.70 (Terwee *et al.* 2007).

COSMIN: no asterisk = indeterminate (grey literature).

*1 = poor.

**2 = fair.

***3 = good.

****4 = excellent.

Version: 24 = ASQ 24-month questionnaire; 27 = ASQ 27-month questionnaire; 30 = ASQ 30-month questionnaire.

Type: ASQ = ASQ-3™; ASQ:T = ASQ-3™, translated; SE = ASQ:SE; SE:T = ASQ:SE, translated.

? No information provided.

n/a = does not.

Reliability and validity of the ASQ:SE

The two ASQ:SE age versions were found to have high values for internal consistency reliability, with Cronbach's alpha values that ranged from 0.71 (Heo 1999) to 0.80 (Squires *et al.* 2001a) for the 24-month version and a value of 0.88 (Squires *et al.* 2001a) for the 30-month version.

The internal consistency reliability of the translated/adapted versions of the ASQ:SE was slightly lower: Cronbach's alpha values ranged from 0.62 (de Wolff *et al.* 2013) to 0.76 (Kucuker *et al.* 2011) for the 24-month version and 0.85 (Kucuker *et al.* 2011) for the 30-month version. However, the studies scored 'poor' for this methodological quality on COSMIN and 'poor' on the quality of the measures' translations.

One study (Heo 1999) reported test–retest reliability for the ASQ:SE, and it only provided information for the 24-month version of the measure. Heo's results (Heo 1999) showed high values for test–retest reliability (correlation = 1). The study was completed on a reasonable sample size, but the methodological quality for this psychometric property was 'poor'.

There were no studies reporting inter-rater reliability of the ASQ:SE or its adapted/translated version. A technical report by the measure's developers demonstrated an overall inter-rater reliability of 0.94 (combining ages from 3 to 66 months) (Squires *et al.* 2001b).

There was evidence of inter-rater reliability for the translated/adapted ASQ:SE version in one study (Kucuker *et al.* 2011), with 0.67 for the 24-month version and 0.80 for the 30-month version. However, the quality of the study was 'poor'.

Sensitivity values for the ASQ:SE were 'positive'. For the 24-month version, a value of 0.71 (Squires *et al.* 2001a) was reported which derived from comparing the measure to the established comparator (Child Behavior Checklist [CBCL] (Achenbach 1992)). For the 30-month version, a value of 0.80 (Squires *et al.* 2001a) was found, also compared with the CBCL. This study was rated as 'fair' on the quality for this psychometric property.

Sensitivity values for the adapted/translated ASQ:SE were less consistent. For the 24-month version, sensitivity value was 0.90 (Kucuker *et al.* 2011) and ranged from 0.66 (de Wolff *et al.* 2013) to 1.0 (Heo & Squires 2012) when compared with the CBCL. The 30-month version presented evidence of a 'positive' value (0.78) (Kucuker *et al.* 2011), which dropped dramatically when compared with the CBCL (0.25) (Heo & Squires 2012). In terms of the methodological quality, the studies varied between 'fair' and 'good' ratings.

Specificity values for the translated/adapted ASQ:SE were not consistent. For the 24-month version, values ranged from 0.93 (Squires *et al.* 2001a) to 0.95 (Heo 1999) when compared with the CBCL, while for the 30-month version, only one value of 0.89 (Squires *et al.* 2001a) was found. However, the quality of one of these studies for this psychometric property was 'fair', with another study rated as 'good'.

The specificity of the adapted ASQ:SE measure was found to have 'positive' values. For the 24-month version, the specificity value was 0.95 (Kucuker *et al.* 2011) and ranged from 0.87 (Heo & Squires 2012) to 0.91 (de Wolff *et al.* 2013) when compared with the CBCL; for the 30-month version the specificity value was 0.74 (Kucuker *et al.* 2011) and 0.80 (Heo 1999) when compared with the CBCL. In terms of the methodological quality, the studies varied between 'fair' and 'good' ratings.

Discussion

The aim of this research was to systematically review international evidence regarding the psychometric properties of the ASQ (ASQ-3^{TM} and ASQ:SE) for use as population outcome indicators at 2.5 years in England. We identified 20 papers meeting the inclusion criteria.

In general, the review showed 'positive' values (Terwee *et al.* 2007) for the measures' psychometric properties: 'positive' values for reliability (*alpha >0.70 or test–retest reliability >0.80 or ICC >0.70*) occurred in 11/18 instances reported (with 4 'intermediate' ratings (*alpha = 0.60-0.70 or test–retest = 0.60-0.80 or ICC = 0.50-0.70*) and 3 'negative' ratings (*alpha < 0.60 or test–retest <0.60 or ICC < 0.50*)), for sensitivity in 13/18 (*>0.70*) instances reported (with 3 'intermediate' ratings (*0.50–0.70*) and 2 'negative' ratings (*<0.50*)), and for specificity in 19/19 (*>0.70*) instances reported.

However, only one study, from the Netherlands, compared the psychometric properties of three questionnaires (ASQ:SE, Brief Infant-Toddler Social and Emotional Assessment–BITSEA, and Brief Instrument Psychological and Pedagogical Problem Inventory–KIPPPI) to detect psychosocial problems in toddlers (de Wolff *et al.* 2013). They found that, at 24 months, BITSEA (Briggs-Gowan *et al.* 2004) discriminated most accurately between children with and without problems (sensitivity = 0.84, specificity = 0.90).

Also, in terms of the sensitivity and specificity levels of the ASQ-3^{TM}, only three studies used the most well-established comparator (BSID) as a comparative instrument, which produced mixed findings (with 8 'positive' and 2 'negative' ratings in original and translated versions). Therefore, no firm conclusions can be made. More of the included studies utilized the most well-established comparator (CBCL) to evaluate the sensitivity and specificity of the ASQ:SE, but the values observed were not homogeneous. Despite 11 (out of 15) 'positive' values on this psychometric property, there were three 'negative' and one 'intermediate' values – some of which were reported in 'good' quality studies (rated specifically for this psychometric property).

To compare the findings of the psychometric properties of the ASQ from this review and other measures of child development, systematic reviews of other measures are needed. Still, the psychometric properties of the ASQ seem comparable to other measures'. For example, the PEDS has shown a sensitivity of 79% and a specificity of 79% for 1 to 3 year olds (Bedford *et al.* 2013). Another study combined screening from a range of healthcare professionals, where no particular assessment tool was used (Chakrabarti & Fombonne 2005); out of the 659 children identified as needing further developmental assessment from professionals' screening, 10% ($n=64$) actually needed further assessment and 90% ($n=595$) were on developmental schedule. Nevertheless, it is important to establish the psychometric properties of the ASQ-3™ and ASQ:SE in an English sample given cultural differences in the understanding of what constitutes developmental delay.

Limitations

Our findings should be considered in the context of their limitations. The overall evidence of the psychometric properties of the measures was limited. Moreover, data were heterogeneous, and, consequently, comparison between studies was challenging. Studies not only varied in sample sizes and sampling procedures (e.g. stratified random samples (Squires *et al.* 2001a), at-risk groups (San Antonio *et al.* 2014)) but also in the contexts/countries in which they were conducted (e.g. North America (San Antonio *et al.* 2014), Brazil (Filgueiras *et al.* 2013), China (Bian *et al.* 2012)). Importantly, differences in study design reflected variations in the aims of the reviewed papers. For approximately half the identified studies, the main aim was to evaluate the measures' psychometric properties (Squires *et al.* 2009b; Bian *et al.* 2012; Filgueiras *et al.* 2013). For the other studies, the psychometric assessments were only part of the subsidiary analyses (Ivey-Soto 2008; Kvestad *et al.* 2013). In addition, some studies employed trained researchers who guided the parent through the assessment (Bian *et al.* 2012; Kvestad *et al.* 2013) at either on-site (Filgueiras *et al.* 2013) or home appointments (Kvestad *et al.* 2013); these may have had an effect on the reported psychometric properties of the measures. Because of the heterogeneous nature of the evidence, the application of the same assessment tool (i.e. COSMIN) may not have been ideal and the scores obtained might not be a true representation of the studies' quality. Moreover, the thresholds used in the COSMIN checklist to classify 'positive' values may be considered low.

The differences in cultural and contextual factors may limit the generalizability of the evidence of the psychometric properties of the ASQ-3™ and ASQ:SE to other dissimilar populations. Not surprisingly, the personal–social and problem-solving sub-scales of the ASQ-3™, which were shown to be the most culture-specific, were also the most affected by the translation/adaptation process, resulting in the lowest Cronbach's alphas. The ASQ and ASQ:SE were translated and adapted in different ways and even with the inclusion of translation quality criteria, the variability of all these different contexts could not be comprehensively gauged.

In terms of the measures' measurement precision (i.e. reliability), it is essential to note that the included studies evaluated this psychometric property using Cronbach's alpha, which makes very strong assumptions of unidimensionality and equal factor loadings. However, these assumptions are almost never tested in applied studies. To ensure the appropriateness of alpha as the index of test reliability, the factorial structure of the instrument must be assessed (Sijtsma & Emons 2011).

Also, the reports on the sensitivity and specificity of the measures may require caution as they depend on how cut-off scores defining 'positivity' were derived and which comparator measure was used, along with its own limitations (McGrath *et al.* 2004; Anderson *et al.* 2010; Moore *et al.* 2012; Spittle *et al.* 2013). Besides, this review only focuses on three age bands, which limits generalisability and significantly reduces the sample sizes used in each study. Thus, differences between the measures, along with their limitations, need to be taken into account when interpreting findings. Finally, the second edition of the ASQ:SE will be published shortly, and its psychometric properties will need study, which may vary from the findings presented here.

Recommendations for future research

Future research should examine the psychometric properties of all age bands as this review. More research is needed to examine the psychometric properties of the measures on an English sample. A range of options are possible, depending on the existing data available and the scope of resources for collecting new data. Particular attention should be paid to the culturally dependent sub-scales in any initial data to explore their reliability. Additionally, the standardization of norms and development of cut-off scores should be conducted in samples drawn from the same population to which they will be applied, with appropriate consideration of relevant demographic characteristics shown to be associated with ASQ scores.

Table 5. Number of children who would be classified as true positive, false negatives, true negatives and false positives if ASQ-3^{TM} and ASQ:SE were to be used in the general population

Prevalence of developmental delay	True positives[a] *N* (%)	False negatives[b] *N* (%)	True negatives[c] *N* (%)	False positives[d] *N* (%)	Total positives *N* (%)	Total negatives *N* (%)
4.5%[e]	348 (3.5%)	104 (1.0%)	7447 (74.5%)	2101 (21.0%)	2449 (24.5%)	7551 (75.5%)

[a] Children with developmental needs who are identified as needing further assessment.
[b] Children with developmental needs who are not identified as needing further assessment.
[c] Children without developmental needs who are identified as being on schedule.
[d] Children without developmental needs who are identified as needing further assessment.
[e] Percentage of children presenting developmental delay (Chakrabarti & Fombonne 2005; Emerson *et al.* 2009)

Implication for practice

The reliability, sensitivity and specificity of the translated/adapted ASQ-3^{TM} and ASQ:SE questionnaires were generally more mixed than the original questionnaires'. This may in part be explained by translation problems; the included studies generally scored 'low' on translation quality. However, there is likely to be variation based on language and also cultural differences, even when comparing between North America and England. These warrant consideration, and there has been some attempt to adapt the measures based on these cultural differences (e.g. current work to adapt the measures for use with English samples (Kendall *et al.* 2014)). Differences in cultural and contextual factors should be considered when measuring child development and determining what would be appropriate for a child at a given age. The personal–social and problem-solving sub-scales showed the lowest levels of reliability when used in non-English speaking countries; these sub-scales refer to culturally dependent behaviours, such as the use of eating tools.

To illustrate the potential implications for practice of using the ASQ-3^{TM} and ASQ:SE questionnaires as population outcome indicators, a worked example was calculated. Calculations were based on the average sensitivity and specificity of the ASQ-3^{TM} and ASQ:SE questionnaires from the review (0.77 and 0.78, respectively). As the prevalence of developmental delay for 2–2.5 years old in England is currently unknown the average of the percentage of 0 to 3.5 years old[3] with developmental delays identified by two previous studies conducted in the UK[4] was used as a proxy (Chakrabarti & Fombonne 2005; Emerson *et al.* 2009). A base-rate of 10 000 2 to 2.5-year-old children was used for the worked example for ease of interpretation. Results are shown in Table 5.

This is not to say that the ASQ-3^{TM} and ASQ:SE present particular issues with respect to accuracy – other measures of child development may not be more precise or valid. Systematic reviews of other measures of child development are needed to compare our findings. For example, the Parent's Evaluation of Developmental Status has a sensitivity of 79% and a specificity of 79% for 1 to 3 year olds (Bedford *et al.* 2013).

Conclusions

This is the first systematic review of the psychometric properties of the current versions of the Ages & Stages Questionnaires relevant to the use of the measures as a population outcome indicator. The findings were generally 'positive' for the reliability, sensitivity and specificity of the original versions of the ASQ-3^{TM} and ASQ:SE. In contrast, the psychometric properties of translated/adapted ASQ-3^{TM} and ASQ:SE questionnaires were more mixed, particularly for more culturally specific domains. This highlights the need for cultural and contextual differences to be considered when measuring child development and determining what would be appropriate for a child at a given age. However, the existing evidence included in this review was generally 'low' quality, meaning that further research is very likely to have an important impact on the interpretation of the ASQ-3^{TM} and ASQ:SE psychometric evidence. Future research is needed to examine the reliability and validity of the measures for an English sample. Training materials may be useful to consider for administering, completing and scoring the questionnaires. Through triangulating measures of child development with other information, such as prospective academic attainment, we may be able to build a picture of

[3] These articles only gave information for the whole age range and therefore, it is not possible to refine this for 2 to 2.5-year-old children.
[4] These two studies were not identified by a systematic review, and so they might be presenting a biassed estimate. However, they were chosen because they were the only two found to report the percentage of children with developmental delay *in the UK*.

the population of children on developmental schedule and those in need of further developmental assessment.

Key messages

- This is the first systematic review of the psychometric properties of the current versions of the Ages & Stages Questionnaires relevant to the use of the measures as a population outcome indicator. The findings were generally positive for the measures' reliability, sensitivity and specificity.
- The reliability, sensitivity, and specificity of the translated/adapted ASQ-3^{TM} and ASQ:SE questionnaires were generally more mixed compared with the original questionnaires, particularly for more culturally specific domains.
- Differences in cultural and contextual factors should be considered when measuring child development and determining what would be appropriate for a child at a given age.

Funding source

The review was funded by the Policy Research Programme in the Department of Health.

Financial disclosure

None.

Conflict of interest

The authors have no conflicts of interest relevant to this article to disclose.

Disclaimer

This is an independent review commissioned and funded by the Policy Research Programme in the Department of Health. The views expressed are not necessarily those of the Department.

Contributors' statements

Dr Velikonja conducted database searches, filtering, conducted quality assessment and interpretation of results and reviewed and revised the manuscript.

Dr. Edbrooke-Childs and Dr. Deighton conceived of the study, contributed to study design, data interpretation and reviewed and revised the manuscript.

Dr. Sleed contributed to data filtering, quality assessments and reviewed the manuscript.

Dr. Brown is an expert on psychometric measurement, contributed to the study design, data interpretation and also reviewed the manuscript.

Dr. Calderon contributed to the review and interpretation of findings, prepared the first draft of the manuscript and reviewed the revised versions.

Acknowledgements

We would like to thank members of the Policy Research Unit in the Health of Children, Young People and Families: Terence Stephenson, Catherine Law, Amanda Edwards, Steve Morris, Helen Roberts, Cathy Street, Russell Viner and Miranda Wolpert.

Special thanks go to the advisory team members: Catherine Law, Miranda Wolpert, Ruth Gilbert, Helen Roberts, Helen Bedford, Steve Pilling and Jane Barlow.

References

Achenbach, T. M. (1992) *Manual for the Child Behavior Checklist/2-3 and 1992*. Profile, Burlington, VA.

Anderson, P. J., De Luca, C. R., Hutchinson E., Roberts, G., Doyle, L. W. & The Victorian Infant Collaborative Group (2010) Underestimation of developmental delay by the new Bayley-III scale. *Archives of Pediatrics and Adolescent Medicine*, **164**, 352–356. DOI:10.1001/archpediatrics.2010.20.

Baggett, K. M., Warlen, L., Hamilton, J. L., Roberts, J. L. & Staker, M. (2007) Screening infant mental health indicators: an early head start initiative. *Infants & Young Children*, **20**, 300–310. DOI:10.1097/01.IYC.0000290353.39793.ba.

Bayley, N. (1993) *The Bayley Scales of Infant Development*. Harcort Brace & Company, New York, London.

Bedford, H., Walton, S., Ahn, J. (2013) *Measures of child development: a review*. Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health.

Bian, X., Yao, G., Squires, J., Hoselton, R., Chen, C.-I., Murphy, K., Wei, M. & Fang, B. (2012) Translation and use of parent-completed developmental screening test in Shanghai. *Journal of Early Childhood Research*, **10**, 162–175. DOI:10.1177/1476718x11430071.

Bian, X., Yao, G., Squires, J., Wei, M., Chen, C. & Fang, B. (2010) Studies of the norm and psychometric properties of the Ages and Stages Questionnaires in Shanghai children. *Chinese Journal of Pediatrics*, **48**, 492–496.

Briggs-Gowan, M. J. & Carter, A. S. (2008) Social–emotional screening status in early childhood predicts elementary school outcomes. *Pediatrics*, **121**, 957–962. DOI:10.1542/peds.2007-1948.

Briggs-Gowan, M. J., Carter, A. S., Irwin, J. R., Wachtel, K. & Cicchetti, D. V. (2004) The brief infant–toddler social and emotional assessment: screening for social–emotional problems and delays in competence. *Journal of Pediatric Psychology*, **29**, 143–155. DOI:10.1093/jpepsy/jsh017.

Brugman, E., Reijneveld, S. A., Verhulst, F. C. & Verloove-Vanhorick, S. (2001) Identification and management of psychosocial problems by preventive child health care. *Archives of Pediatrics and Adolescent Medicine*, **155**, 462–469. DOI:10.1001/archpedi.155.4.462.

Chakrabarti, S. & Fombonne, E. (2005) Pervasive developmental disorders in preschool children: confirmation of high prevalence. *American Journal of Psychiatry*, **162**, 1133–1141. DOI:10.1176/appi.ajp.162.6.1133.

Cosmin Group. Systematic reviews of measurement properties. Available at: www.cosmin.nl (last accessed 5 June 2014)

Council on Children with Disabilities, Section on Developmental Behavioral Pediatrics, Bright Futures Steering Committee & Medical Home Initiatives for Children with Special Needs Project Advisory Committee (2006) Identifying infants and young children with developmental disorders in the medical home: an algorithm for developmental surveillance and screening. *Pediatrics*, **118**, 405–420. DOI:10.1542/peds.2006-1231.

D'Aprano, A., Silburn, S., Johnston, V., Robinson, G., Oberklaid, F. & Squires, J. (2014) Adaptation of the Ages and Stages Questionnaire for remote Aboriginal Australia. *Qualitative Health Research*. DOI:10.1177/1049732314562891.

de Wolff, M. S., Theunissen, M. H. C., Vogels, A. G. C. & Reijneveld, S. A. (2013) Three questionnaires to detect psychosocial problems in toddlers: a comparison of the BITSEA, ASQ:SE, and KIPPPI. *Academic Pediatrics*, **13**, 587–592. DOI:10.1016/j.acap.2013.07.007.

Department of Health (2014) *Factsheet: Developing a Public Health Outcome Measure for Children Aged 2-2½ Using ASQ-3TM*. Department of Health, London.

Eeles, A. L., Spittle, A. J., Anderson, P. J., Brown, N., Lee, K. J., Boyd, R. N. & Doyle, L. W. (2013) Assessments of sensory processing in infants: a systematic review. *Developmental Medicine and Child Neurology*, **55**, 314–326. DOI:10.1111/j.1469-8749.2012.04434.x.

Emerson, E., Graham, H., Mcculloch, A., Blacher, J., Hatton, C. & Llewellyn, G. (2009) The social context of parenting 3-year-old children with developmental delay in the UK. *Child: Care, Health and Development*, **35**, 63–70. DOI:10.1111/j.1365-2214.2008.00909.x.

Feil, E. G., Severson, H. H. & Walker, H. M. (1998) Screening for emotional and behavioral delays: the early screening project. *Journal of Early Intervention*, **21**, 252–266. DOI:10.1177/105381519802100306.

Field, D. & Livingstone, R. (2013) Clinical tools that measure sitting posture, seated postural control or functional abilities in children with motor impairments: a systematic review. *Clinical Rehabilitation*, **27**, 994–1004. DOI:10.1177/0269215513488122.

Filgueiras, A. (2014) *Inter-Temporal Stability of the ASQ:BR Using the Teacher's Responses*. Poster presentation, ASQ Around the World Symposium, San Francisco.

Filgueiras, A., Pires, P., Maissonette, S. & Landeira-Fernandez, J. (2013) Psychometric properties of the Brazilian-adapted version of the Ages and Stages Questionnaire in public child daycare centers. *Early Human Development*, **89**, 561–576. DOI:10.1016/j.earlhumdev.2013.02.005.

Flamant, C., Branger, B., Nguyen The Tich, S., de La Rochebrochard, E., Savagner, C., Berlie, I. & Rozé, J.-C. (2011) Parent-completed developmental screening in premature children: a valid tool for follow-up programs. *PLoS ONE*, **6**, e20004. doi: 10.1371/journal.pone.0020004

Frankenburg, W. K., Doods, J., Archers, P., Shapior, H. & Bresnick, B. (1990) *DDST-II: Denver Developmental Screening Test*. Denver Developmental Matierals, Inc., Denver.

Heo, G. (1999) Early identification of social–emotional competence in young children: a study of the Ages and Stages Questionnaires: Social–Emotional (ASQ:SE). University of Oregon.

Heo, K. H. & Squires, J. (2012) Cultural adaptation of a parent completed social emotional screening instrument for young children: ages and stages questionnaire-social emotional. *Early Human Development*, **88**, 151–158. DOI:10.1016/j.earlhumdev.2011.07.019.

Higgins, J. P. T. & Green, S. (2011) *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0*. [updated March 2011]. The Cochrane Collaboration. Available from http://handbook.cochrane.org

Ivey-Soto, M. C. (2008) Examining the utility of a new caregiver-completed social emotional assessment, the Social Emotional Assessment Measure, with diverse low-income-toddler dyads. University of Oregon.

Juneja, M., Mohanty, M., Jain, R. & Ramji, S. (2012) Ages and Stages Questionnaire as a screening tool for developmental delay in Indian children. *Indian Pediatrics*, **49**, 457–461.

Kendall, S., Nash, A., Brown, A., Bastug, G., Rougeux, E., Bedford, H. (2014) *Evaluating the use of a population measure of child development in the Healthy Child Programme Two year review*. Centre for Paediatric Epidemiology and Biostatistics, UCL Institute of Child Health and School of Health and Social Work University of Hertfordshire.

Kerstjens, J. M., Bos, A. F., Ten Vergert, E. M. J., De Meer, G., Butcher, P. R. & Reijneveld, S. A. (2009) Support for the global feasibility of the Ages and Stages Questionnaire as developmental screener. *Early Human Development*, **85**, 443–447. DOI:10.1016/j.earlhumdev.2009.03.001.

Kucuker, S., Kapci, E. G. & Uslu, R. I. (2011) Evaluation of the Turkish version of the 'Ages and Stages Questionnaires: Social–Emotional' in identifying children with social-emotional problems. *Infants and Young Children*, **24**, 207–220.

Kvestad, I., Taneja, S., Kumar, T., Bhandari, N., Strand, T. A. & Hysing, M. (2013) The assessment of developmental status using the Ages and Stages questionnaire-3 in nutritional research in north Indian young children. *Nutrition Journal*, **12**, 1–11. DOI:10.1186/1475-2891-12-50.

Landis, J. R. & Koch, G. G. (1977) The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174. DOI:10.2307/2529310.

Lopes, S., Serrano, A. M., Teixeira, S. & Graca, P. (2014) *Portuguese Version of the Ages & Stages Questionnaire: The Results of the Validation/Standardization of the ASQ-3 for Portuguese Population*. Oral presentation, ASQ Around the World Symposium, San Francisco.

Luttikhuizen Dos Santos, E. S., De Kieviet, J. F., Königs, M., Van Elburg, R. M. & Oosterlaan, J. (2013) Predictive value of the Bayley Scales of Infant Development on development of very preterm/very low birth weight children: a meta-analysis. *Early Human Development*, **89**, 487–496. DOI:10.1016/j.earlhumdev.2013.03.008.

McCullough, N. & Parkes, J. (2008) Use of the child health questionnaire in children with cerebral palsy: a systematic review and evaluation of the psychometric properties. *Journal of Pediatric Psychology*, **33**, 80–90. DOI:10.1093/jpepsy/jsm070.

McGrath, E., Wypij, D., Rappaport, L. A., Newburger, J. W. & Bellinger, D. (2004) Prediction of IQ and achievement at age 8 years from neurodevelopmental status at age 1 year in children with D-transposition of the great arteries. *Pediatrics*, **114**, e572–e576. DOI:10.1542/peds.2003-0983-L.

Mick, E., Biederman, J., Pandina, G. & Faraone, S. V. (2003) A preliminary meta-analysis of the child behavior checklist in pediatric bipolar disorder. *Biological Psychiatry*, **53**, 1021–1027. DOI:10.1016/s0006-3223(03)00234-8.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. & The Prisma Group (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine*, **6** e1000097. DOI:10.7326/0003-4819-151-4-200908180-00135.

Moore, T., Johnson, S., Haider, S., Hennessy, E. & Marlow, N. (2012) Relationship between test scores using the second and third editions of the Bayley scales in extremely preterm children. *The Journal of Pediatrics*, **160**, 553–558. DOI:10.1016/j.jpeds.2011.09.047.

Østergaard, K. K., Lando, A. V., Hansen, B. M. & Greisen, G. (2012) A Danish reference chart for assessment of psychomotor development based on the Ages & Stages Questionnaire. *Danish Medical Journal*, **59**, A4429.

Pinto-Martin, J. A., Dunkle, M., Earls, M., Fliedner, D. & Landes, C. (2005) Developmental stages of developmental screening: steps to implementation of a successful program. *American Journal of Public Health*, **95**, 1928–1932. DOI:10.2105/ajph.2004.052167.

Pomes, M. (2013) Examination of the Spanish translation of a developmental screening instrument. University of Oregon.

Saihong, P. (2010) Use of screening instrument in northeast Thai early childcare settings. *Procedia – Social and Behavioral Sciences*, **7**, 97–105. DOI:10.1016/j.sbspro.2010.10.015.

Saihong, P. (2014) *A Translation and Adaptation of the ASQ-3 for Thailand*. Poster presentation, ASQ Around the World Symposium, San Francisco.

San Antonio, M. C., Fenick, A. M., Shabanova, V., Leventhal, J. M. & Weitzman, C. C. (2014) Developmental screening using the ages and stages questionnaire: standardized versus real-world conditions. *Infants & Young Children*, **27**, 111–119. DOI:10.1097/iyc.0000000000000005.

Sarmiento Campos, J. A., Squires, J. & Ponte, J. (2011) Universal developmental screening: preliminary studies in Galicia, Spain. *Early Child Development and Care*, **181**, 475–485. DOI:10.1080/03004430903458007.

Schellingerhout, J. M., Heymans, M. W., Verhagen, A. P., De Vet, H. C., Koes, B. W. & Terwee, C. B. (2011) Measurement properties of translated versions of neck-specific questionnaires: a systematic review. *BMC Medical Research Methodology*, **11**, 1–14. DOI:10.1186/1471-2288-11-87.

Schonhaut, L., Armijo, I., Schönstedt, J., Alvarez, J. & Cordero, M. (2013) Validity of the Ages and Stages Questionnaires in term and preterm infants. *Pediatrics*, **131**, e1468–e1474. DOI:10.1542/peds.2012-3313.

Sheldrick, R. C. & Perrin, E. C. (2013) Evidence-based milestones for surveillance of cognitive, language, and motor development. *Academic Pediatrics*, **13**, 577–586. DOI:10.1016/j.acap.2013.07.001.

Sidor, A., Fischer, C., Eickhorst, A. & Cierpka, M. (2013) Influence of early regulatory problems in infants on their development at 12 months: a longitudinal study in a high-risk sample. *Child and Adolescent Psychiatry and Mental Health*, **7**, 1–14. DOI:10.1186/1753-2000-7-35.

Sijtsma, K. & Emons, W. H. M. (2011) Advice on total-score reliability issues in psychosomatic measurement. *Journal of Psychosomatic Research*, **70**, 565–572. DOI:10.1016/j.jpsychores.2010.11.002.

Spittle, A. J., Spencer-Smith, M. M., Eeles, A. L., Lee, K. J., Lorefice, L. E., Anderson, P. J. & Doyle, L. W. (2013) Does the Bayley-III Motor Scale at 2 years predict motor outcome at 4 years in very preterm children? *Developmental Medicine and Child Neurology*, **55**, 448–452. DOI:10.1111/dmcn.12049.

Squires, J., Bricker, D., Heo, K. & Twombly, E. (2001a) Identification of social–emotional problems in young children using a parent-completed screening measure. *Early Childhood Research Quarterly*, **16**, 405–419. DOI:10.1016/S0885-2006(01)00115-6.

Squires, J., Bricker, D., Twombly, E., Potter, L. (2009a) ASQ technical report. Available at: http://www.agesandstages.com (last accessed 26th June 2014).

Squires, J., Potter, L., Bricker, D. (2001b) Technical report on ASQ:SE.

Squires, J., Twombly, E. & Bricker, D. (2009b) *ASQ-3 User's Guide*. Brookes Publishing Co, Baltimore.

Terwee, C. B., Bot, S. D., De Boer, M. R., Van Der Windt, D. A., Knol, D. L., Dekker, J., Bouter, L. M. & De Vet, H. C. (2007) Quality criteria were proposed for measurement properties of health status questionnaires. *Journal of Clinical Epidemiology*, **60**, 34–42. DOI:10.1016/j.jclinepi.2006.03.012.

Troude, P., Squires, J., L'hélias, L. F., Bouyer, J. & La Rochebrochard, E. D. (2011) Ages and stages questionnaires: feasibility of postal surveys for child follow-up. *Early Human Development*, **87**, 671–676. DOI:10.1016/j.earlhumdev.2011.05.007.

Veldhuizen, S., Clinton, J., Rodriguez, C., Wade, T. J. & Cairney, J. (2014) Concurrent validity of the ages and stages questionnaires and Bayley developmental scales in a general population sample. *Academic Pediatrics*, **15**, 231–237. DOI:10.1016/j.acap.2014.08.002.

Verhulst, F. C., Koot, H. M. & Van Der Ende, J. (1994) Differential predictive value of parents' and teachers' reports of children's problem behaviors: a longitudinal study. *Journal of Abnormal Child Psychology*, **22**, 531–546.

Appendix 1

Search terms

The following two sets of search terms were used: (1) 'Ages and Stages Questionnaire' or 'Age & Stage Questionnaire*' or 'Ages & Stages Questionnaire*' or 'ASQ*' and (2) 'valid*' or 'reliab*' or 'psychometric*' or 'reproducib*' or 'internal consistency' or 'ceiling effect' or 'floor effect' or 'coefficient of variation' or 'discriminative' or 'precision' or 'testing' or 'measurement' or 'applicab*' or 'utility' or 'screening' or 'statistical analysis' or 'test construction' or 'test standardi?ation' or 'test interpretation' or 'reproducibility of results' or 'methods' or 'observer variation' or 'measurement invariance' or 'measurement equivalence' or 'test homogeneity' or 'construct bias'.

Adapted methodological criteria for the translation process and cross-cultural validation only included items 4 to 11 of the original criteria.

1 Was the percentage of missing items given?
2 Was there a description of how missing items were handled?
3 Was the sample size included in the analysis adequate?
4 Were both the original language in which the HR-PRO instrument was developed, and the language in which the HR-PRO instrument was translated described?
5 Was the expertise of the people involved in the translation process adequately described? e.g. expertise in the disease (s) involved, in the construct to be measured, or in both languages
6 Did the translators work independently from each other?
7 Were items translated forward and backward?
8 Was there an adequate description of how differences between the original and translated versions were resolved?
9 Was the translation reviewed by a committee (e.g. original developers)?
10 Was the HR-PRO instrument pre-tested (e.g. cognitive interviews) to check interpretation, cultural relevance of the translation, and ease of comprehension?
11 Was the sample used in the pre-test adequately described?
12 Were the samples similar for all characteristics except language and/or cultural background?
13 Were there any important flaws in the design or methods of the study?
14 for CTT: Was confirmatory factor analysis performed?
15 for IRT: Was differential item function (DIF) between language groups assessed?