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.

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™ 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™ 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™) was published in 2009. The ASQ-3™ 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; Östergaard 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™ (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™ 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 1995¹ (the year questionnaires were first published) and 15 December 2014. All language versions of the ASQ-3™ 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-3TM 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-3TM (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-3TM 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-3TM 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-3TM 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)

© 2016 John Wiley & Sons Ltd, Child: care, health and development, 43, 1, 1–17
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...
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>
<thead>
<tr>
<th>First author, year</th>
<th>Country</th>
<th>Study design</th>
<th>N</th>
<th>Questionnaires version</th>
<th>Age version</th>
<th>Aim(s) of the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Antonio et al. (2014)</td>
<td>USA</td>
<td>Convenience sample, at-risk groups</td>
<td>131</td>
<td>ASQ, English &amp; Spanish</td>
<td>30 (n = 65)</td>
<td>To examine the reproducibility of the ASQ-3&lt;sup&gt;TM&lt;/sup&gt; under standardized versus non-standardized conditions</td>
</tr>
<tr>
<td>Veldhuizen et al. (2014)</td>
<td>Canada</td>
<td>Convenience sample from community organizations</td>
<td>587</td>
<td>ASQ, English</td>
<td>24 (n = 64) 30 (n = 49)</td>
<td>To evaluate the agreement between the ASQ-3&lt;sup&gt;TM&lt;/sup&gt; and the Bayley-II&lt;sup&gt;P&lt;/sup&gt;</td>
</tr>
<tr>
<td>Heo (1999)</td>
<td>USA</td>
<td>Stratified sample</td>
<td>447</td>
<td>SE, English</td>
<td>24 (n = 237)</td>
<td>To examine the validity and reliability of the ASQ:SE</td>
</tr>
<tr>
<td>Ivey-Soto (2008)&lt;sup&gt;†&lt;/sup&gt;</td>
<td>USA</td>
<td>Non-representative sample (Early Head Start programs attendees)</td>
<td>50</td>
<td>SE, English</td>
<td>24 (n = 16) 30 (n = 13)</td>
<td>To assess the utility and usability of the SEAM&lt;sup&gt;a&lt;/sup&gt; To evaluate the relationship between the SEAM and ASQ:SE</td>
</tr>
<tr>
<td>Bian et al. (2012)</td>
<td>China</td>
<td>Cross-sectional study (stratified sampling)</td>
<td>8472</td>
<td>ASQ, Chinese</td>
<td>24, 30</td>
<td>To evaluate the reliability and validity of the Chinese translation of the ASQ-3&lt;sup&gt;TM&lt;/sup&gt; to identify developmental delays in preschool children</td>
</tr>
<tr>
<td>Filgueiras et al. (2013)</td>
<td>Brazil</td>
<td>All public day centres in Rio de Janeiro</td>
<td>45640</td>
<td>ASQ, Portuguese (Brazilian)</td>
<td>24 (n = 1454) 27 (n = 2222) 30 (n = 2814)</td>
<td>To translate and adapt the ASQ-3&lt;sup&gt;TM&lt;/sup&gt; for use in Brazilian public child day care centres To explore the psychometric properties of the Brazilian ASQ-3&lt;sup&gt;TM&lt;/sup&gt; ‘home procedure’ in epidemiological studies</td>
</tr>
<tr>
<td>Kvestad et al. (2013)</td>
<td>India</td>
<td>Sampled part of a randomized double blind placebo control study (last 440 enrolled children – out of 1000 in total – included in this study)</td>
<td>422</td>
<td>ASQ, Indian</td>
<td>24 (n = 39) 27 (n = 47) 30 (n = 37)</td>
<td>To assess the feasibility of the ASQ-3&lt;sup&gt;TM&lt;/sup&gt;</td>
</tr>
<tr>
<td>Pomes (2013)&lt;sup&gt;‡&lt;/sup&gt;</td>
<td>USA</td>
<td>Convenience sample</td>
<td>798</td>
<td>ASQ, Spanish (Chilean)</td>
<td>30 (n = 177 Spanish and n = 127 English)</td>
<td>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&lt;sup&gt;TM&lt;/sup&gt;</td>
</tr>
<tr>
<td>Schonhaut et al. (2013)</td>
<td>Chile</td>
<td>Representative sample, preterm and term children, recruited from ambulatory well-child clinic</td>
<td>306</td>
<td>ASQ, Spanish (Chilean)</td>
<td>30 (n = 96)</td>
<td>To assess the concurrent validity of the ASQ-3&lt;sup&gt;TM&lt;/sup&gt; Chile – compared with the Bayley Scales of Infant and Toddler Development (Bayley-III)</td>
</tr>
<tr>
<td>de Wolff et al. (2013)</td>
<td>Netherlands</td>
<td>Cross-sectional representative sample</td>
<td>2106</td>
<td>SE, Dutch</td>
<td>24 (n = 840)</td>
<td>To compare the psychometric properties of the three questionnaires to detect psychosocial problems in toddlers</td>
</tr>
<tr>
<td>Heo and Squires (2012)</td>
<td>Korea</td>
<td>Stratified sample (Korean census data)</td>
<td>2562</td>
<td>SE, Korean</td>
<td>24 (n = 293) 30 (n = 206)</td>
<td>To investigate the appropriateness of the translation of the Korean ASQ:SE and its validity and reliability</td>
</tr>
</tbody>
</table>

(continues)
### Table 1. (Continued)

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

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.

* The Social Emotional Assessment Measure (SEAM).  
† 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

<table>
<thead>
<tr>
<th>First author, year</th>
<th>Country</th>
<th>Study design</th>
<th>N</th>
<th>Questionnaires version</th>
<th>Age version</th>
<th>Aim(s) of the study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squires et al. (2001a, 2009a)</td>
<td>USA</td>
<td>Stratified sample of US population</td>
<td>18572</td>
<td>ASQ, English</td>
<td>24, 27, 30</td>
<td>To assess the validity and reliability of the ASQ-3™</td>
</tr>
<tr>
<td>Squires et al. (2001b)</td>
<td>USA</td>
<td>Stratified sample of US population</td>
<td>3014</td>
<td>SE, English</td>
<td>24, 30</td>
<td>To assess the validity and reliability of the ASQ:SE</td>
</tr>
<tr>
<td>Filgueiras (2014)</td>
<td>Brazil</td>
<td>Convenience sample</td>
<td>150</td>
<td>ASQ, Portuguese (Brazilian)</td>
<td>24 (n = 50) 27 (n = 50) 30 (n = 50)</td>
<td>To assess the test-retest reliability of the Brazilian ASQ-3™ using teachers’ responses</td>
</tr>
<tr>
<td>Lopes et al. (2014)</td>
<td>Portugal</td>
<td>General population sample</td>
<td>1908</td>
<td>ASQ, Portuguese</td>
<td>24 (n = 111) 27 (n = 109) 30 (n = 112)</td>
<td>Evaluate the psychometric properties of the ASQ-3™ Portuguese</td>
</tr>
<tr>
<td>Saihong (2014)</td>
<td>Thailand</td>
<td>Convenience sample</td>
<td>60</td>
<td>ASQ, Thai</td>
<td>24 (n = 30) 30 (n = 30)</td>
<td>To assess the concurrent validity of the Thai ASQ-3™ compared with the Denver Development Screening II (DOST-II)</td>
</tr>
</tbody>
</table>

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™ and ASQ:SE and their translated/adapted versions

<table>
<thead>
<tr>
<th>First author, year</th>
<th>Country</th>
<th>Type</th>
<th>Internal consistency</th>
<th>Reliability indicators</th>
<th>Validity indicators</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>San Antonio et al. (2014)</td>
<td>USA</td>
<td>ASQ</td>
<td>24</td>
<td>27</td>
<td>30</td>
<td>24</td>
<td>27</td>
</tr>
<tr>
<td>Heo (1999)</td>
<td>USA</td>
<td>SE</td>
<td>0.71*</td>
<td>n/a</td>
<td>?</td>
<td>1*</td>
<td>n/a</td>
</tr>
<tr>
<td>Kucuker et al. (2011)</td>
<td>Turkey</td>
<td>SE: T</td>
<td>0.76*</td>
<td>n/a</td>
<td>?</td>
<td>0.85*</td>
<td>0.67*</td>
</tr>
</tbody>
</table>

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; SET = 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™ 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™ 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™ (see Table 4).

The internal consistency reliability of the translated/adapted versions of the ASQ-3™ 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™, 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™ 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™ and only one unpublished study (Lopes et al. 2014) that examined the inter-rater reliability of the translated/adapted ASQ-3™. 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™ 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™ 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 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™ 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™ 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>
<thead>
<tr>
<th>First author, year</th>
<th>Country</th>
<th>Type</th>
<th>Internal consistency</th>
<th>Test-retest</th>
<th>Inter-rater</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Squires et al. (2009a)</td>
<td>USA</td>
<td>ASQ</td>
<td>0.77</td>
<td>0.78</td>
<td>0.78</td>
<td>0.91</td>
<td>0.78</td>
</tr>
<tr>
<td>Squires et al. (2001b)</td>
<td>USA</td>
<td>SE</td>
<td>0.80*</td>
<td>n/a</td>
<td>0.88*</td>
<td>0.71</td>
<td>n/a</td>
</tr>
<tr>
<td>Filgueiras (2014)</td>
<td>Brazil</td>
<td>ASQ:T</td>
<td>?</td>
<td>?</td>
<td>0.73</td>
<td>0.73</td>
<td>0.71</td>
</tr>
<tr>
<td>Lopes et al. (2014)</td>
<td>Portugal</td>
<td>ASQ:T</td>
<td>0.46</td>
<td>0.57</td>
<td>0.52</td>
<td>0.94</td>
<td>0.89</td>
</tr>
</tbody>
</table>

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; SET = 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 measure’s 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™ 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™, 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 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™ and ASQ:SE were to be used in the general population

<table>
<thead>
<tr>
<th>Prevalence of developmental delay</th>
<th>True positives* N (%)</th>
<th>False negativesb N (%)</th>
<th>True negativesc N (%)</th>
<th>False positivesd N (%)</th>
<th>Total positives N (%)</th>
<th>Total negatives N (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.5%*</td>
<td>348 (3.5%)</td>
<td>104 (1.0%)</td>
<td>7447 (74.5%)</td>
<td>2101 (21.0%)</td>
<td>2449 (24.5%)</td>
<td>7551 (75.5%)</td>
</tr>
</tbody>
</table>

* 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.

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™ 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™ 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™ and ASQ:SE questionnaires from the review (0.77 and 0.78, respectively). As the prevalence of developmental delay for 2–2.5-year-old children in England is not possible to refine this for 2 to 2.5-year-old children, 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™ 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™ and ASQ:SE. In contrast, the psychometric properties of translated/adapted ASQ-3™ 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™ 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
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™ 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**


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


**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 standardization’ 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?