Research in Developmental Disabilities 34 (2013) 3267–3275

Contents lists available at SciVerse ScienceDirect

Research in Developmental Disabilities

The Mandarin Chinese version of the childhood autism spectrum test (CAST): Test–retest reliability☆

Xiang Sun [a,b,*], Carrie Allison [b], Bonnie Auyeung [b], Fiona E. Matthews [c], Simon Baron-Cohen [b], Carol Brayne [a]

[a] Cambridge Institute of Public Health, Department of Public Health and Primary Care, Forvie Site, Robinson Way, University of Cambridge, CB2 0SR, UK
[b] Autism Research Centre, Department of Psychiatry, Douglas House, 18b Trumpington Road, University of Cambridge, CB2 2AH, UK
[c] MRC Biostatistics Unit, Cambridge Institute of Public Health, Forvie Site, Robinson Way, University of Cambridge, CB2 0SR, UK

ARTICLE INFO

Article history:
Received 2 March 2013
Received in revised form 23 May 2013
Accepted 23 May 2013
Available online

Keywords:
Autism
Screening
Test–retest reliability
CAST
China

ABSTRACT

This study aimed to investigate the test–retest reliability of a Mandarin Chinese version of the Childhood Autism Spectrum Test (CAST), in a Chinese population. Parents in a school based study on the prevalence of ASC in mainland China were asked to complete a second version of the CAST approximately 2–4 months after first completion. Test retest data were available from 70 children (questionnaires completed by the same parent). Using a cut-off score of 15, the test–retest reliability was good (kappa = 0.64). The test–retest reliability in three categories ($\leq$11, 12–14, $\geq$15) was moderate (weighted kappa = 0.53). The correlation between the scores at CAST-1 and CAST-2 was good (Spearman rho = 0.73). The Mandarin CAST demonstrated moderate to good test–retest reliability as a screening instrument for ASC in an assessment sample in mainland China.

1. Introduction

Autism Spectrum Conditions (ASC) are characterised by impairments in social interaction, communication, alongside the presence of usually repetitive and stereotyped behaviours, and usually narrow interests and activities (American Psychiatric Association, 2000). ASC are conceptualised to lie on a continuum, with degrees of severity, ranging from individuals diagnosed with childhood autism to milder manifestations of the condition such as Asperger Syndrome (Chlebowski, Green, Barton, & Fein, 2010). As a neurodevelopmental condition, the impairments associated with ASC persist across the lifespan (World Health Organisation, 1993). Targeted intervention may reduce the risk of secondary difficulties and help to improve the quality of life of people with ASC (Bryson, Rogers, & Fombonne, 2003; Dover & Le, 2007). Thus, early detection of ASC is necessary to increase the value of early appropriate interventions (Baron-Cohen et al., 2000; Vostanis, Smith, Chung, & Corbett, 1994).

Recent epidemiological studies in the West report the prevalence of ASC to be around 1% of the general population (Baron-Cohen et al., 2009). These studies have adopted a two-phase process for case identification including: screening followed by diagnostic assessment, and using standardised diagnostic instruments (Fombonne, 2009). Outside the West, less

* Corresponding author at: Cambridge Institute of Public Health, Department of Public Health and Primary Care, Forvie Site, Robinson Way, University of Cambridge, CB2 0SR, UK. Tel.: +44 01223 763833; fax: +44 01223 330300.
E-mail address: xs227@medschl.cam.ac.uk (X. Sun).

0891-4222/$ – see front matter © 2013 The Authors. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ridd.2013.05.042

research has been conducted on ASC. To date, little is known about the prevalence of ASC in China. An early review suggested the prevalence of ASC in mainland China is around 0.1%, much lower than Western estimates (Sun & Allison, 2009). However, research methodologies in existing studies are different from those in the West, which has led to difficulties in comparing results (Tang, Guo, Rice, Wang, & Cubells, 2010; Zhang & Ji, 2005). One major difference is the choice of screening and diagnostic instruments. Mainland China has a large population and most screening tools that have been used were developed from the West more than three decades ago (Wang, Wang, & Wang, 2003; Yang, Huang, Jia, & Chen, 1993).

Potential screening instruments for ASC should be evaluated in a Chinese population. When examining the utility of a screening instrument for ASC, both validity and reliability need be determined. Reliability can be examined by conducting a test–retest study. The test–retest reliability of screening instruments for ASC has been reported in previous studies using varying approaches (see Table 1). The reliability of the Childhood Autism Rating Scale (CARS) was first investigated by examining the inter-rater agreement for a particular cut-off using Cohen's kappa (Cohen, 1960; Pereira, Riesgo, & Wagner, 2008), and by calculating the descriptive statistic, the intra-class correlation, to assess the consistency of quantitative measurements made by observers (Yen & Lo, 2002). Cohen's kappa was also used in a reliability study of the Gilliam Autism Disorder Scale (GADS) (Gilliam, 2003). The test–retest reliability of the Autism Behaviour Checklist (ABC) (Goodman & Minne, 1995), the Social Communication Questionnaire (SCQ) (Gau et al., 2011) and the Social Responsiveness Scale (SRS) (Bolte, Poustka, & Constantino, 2008) was examined by calculating the intra-class correlation coefficient. The SCQ (Bolte, Crecelius, & Poustka, 2000), the ASSQ (Ehlers & Gillberg, 1993) and the SRS (Pine, Luby, Abbacchi, & Constantino, 2006) were also analysed using a Pearson's correlation coefficient. One study of the ASSQ used a paired t-test to detect whether the disagreement between the test pairs was random (Ehlers, Gillberg, & Wing, 1999).

The Childhood Autism Spectrum Test (CAST) is a screening instrument developed in the UK to detect ASC in children aged 4–11 years old. It was developed specifically for children in mainstream schools to detect autistic behaviours because many children with ASC (especially those with subtle manifestations of the condition such as Asperger syndrome) are often not identified before primary school (Kamio, 2007; Williams, 2003). The CAST was previous known as the Childhood Asperger Screening Test (Scott, Baron-Cohen, Bolton, & Brayne, 2002). The same acronym and the items were retained, but the title was modified since the same instrument can be used to detect all ASC, not just Asperger Syndrome (Baron-Cohen et al., 2009). Previous studies have provided evidence that the CAST can be used as a screening instrument in large population-based epidemiological research for ASC. The recommended cut-off on the CAST is 15. The sensitivity of CAST at that cut-point in a population setting was 100%, specificity was 97%, and positive predictive value (PPV) was 50% (Williams et al., 2005). The CAST was used as a screening instrument to detect children with possible ASC in a large prevalence study of children aged 5–9 in mainstream schools in Cambridgeshire, which reported a prevalence estimate of 157 per 10,000 in the UK (Baron-Cohen et al., 2009).

The test–retest reliability of the UK CAST was found to be good when it was examined across two score groups (<15 versus ≥ 15) in a population sample of children in mainstream schools (kappa = 0.70) (Williams et al., 2006). Another study examined the test–retest reliability of the CAST in a sample enriched with high scorers in the UK (Allison et al., 2007) in order to see how reliable the instrument was across the threshold of 15. Moderate agreement was reported (kappa = 0.41) across two score groups (<15 versus ≥ 15) and 73.8% children did not move between score groups. This study also found fair test–retest reliability across three score groups (≤ 11, 12–14, ≥ 15) (kappa = 0.25).

In order to use the CAST in large population-based epidemiological studies in China, it is important to examine its test–retest reliability. Another study has examined the validity of the CAST in a Chinese mainstream school population and will be reported elsewhere (Sun et al., 2012). This study examined the test retest reliability of the Mandarin version of the CAST in mainland China in a high scoring sample, using the same methodology reported in Allison et al. (2007).

2. Methods

2.1. Participants

The CAST (CAST-1) was distributed to the parents of 737 children aged 6–11 years (school years 1–4) in two mainstream schools in Beijing city. The questionnaires were distributed by school teachers for parental completion at home. The returned questionnaires were collected from teachers. In total, 714 questionnaires (response rate: 97%, not unusual in Chinese studies) were returned. Those children who scored at or above 15 (≥ 15), and those who scored between 12 and 14 (12–14), were invited to a diagnostic assessment, as were 5% of randomly selected children who scored less than 12. The parents were informed that an invitation for further assessment did not mean that their children had problems. The researchers explained to them that this research selected children in all score groups in order to get more representative presentation of social and communication behaviours of the children in mainstream schools. The responders were contacted two months after the distribution of CAST-1 and invited for a detailed diagnostic assessment. The CAST (CAST-2) was distributed again to participants who came for the diagnostic assessment. The diagnostic assessment included the Autism Diagnostic Observation Schedule (ADOS) (Lord, Rutter, DiLavore, & Risi, 2001) and the Autism Diagnostic Interview-Revised (ADI-R) (Rutter, LeCouteur, & Lord, 2003). The CAST-2 was administered before the diagnostic assessments. The time lag between the distribution of the CAST-1 and CAST-2 was 2–4 months. In order to encourage participation, the parents were informed by an invitation letter about the purpose of this study was to examine the social and communication ability of the child. The

Table 1
Reported test–retest reliability of screening instruments for ASC in primary school aged children.

Screening instrument	Candidate	Year	Sample size and age	Sample source	Time interval	Method	Result
Childhood Autism Rating Scale (CARS)	Pereira (Pereira et al., 2008)	2008	50 with autism: 3–17 years old	University hospital patients	Minimum 4 weeks	Kappa statistic	$r = 0.90$
	Russell (Russell, Kelly, & Golding, 2010)	2010	103 with autism: 22–44.5 years old	Clinical patients	12 months	Intra class correlation coefficient	ICC = 0.81
Autism Behaviour Checklist (ABC)	Goodman (Goodman & Minne, 1995)	1995	17 blind children: 4–11 years old	Clinical patients	11 weeks	Intra class correlation coefficient	ICC = 0.65 for teachers; ICC = 0.21 for parents
Gilliam Asperger's Disorder Scale (GADS)	Gilliam (Gilliam, 2003)	2003	468	N/A	2 weeks	Kappa statistic	$r = 0.93$
Autism Spectrum Screening Questionnaire (ASSQ)	Ehlers (Ehlers & Gillberg, 1993)	1993	139: 7–16 years old	Epidemiological study sample	8 months	Pearson's correlation coefficient	$r = 0.90$
	Ehlers (Ehlers et al., 1999)	1999	65 (teacher version); 86 (parent version): 6–17	Clinical patients	2 weeks	Pearson's correlation coefficient	$r = 0.94$ (teacher) $r = 0.96$ (parent)
Social Communication Questionnaire (SCQ)	Gau (Gau et al., 2011)	2011	86 with ASC: 2–18 years old	Clinical patients	2 weeks	Intra class correlation coefficient	ICC = 0.77–0.78
	Bolte (Bolte et al., 2000)	2000	17 with ASC	N/A	12–24 months	Pearson's correlation coefficient	$r = 0.74$
Social Responsiveness Scale (SRS)	Pine (Pine et al., 2006)	2006	22 with ASC; 51 normal	Preschool children	1 month	Pearson's correlation coefficient	$r = 0.75$
	Bolte (Bolte et al., 2008)	2008	838 normal	Preschool children	3–6 months	Intra class correlation coefficient	ICC = 0.84–0.97
Childhood Autism Spectrum Test (CAST)	Williams (Williams et al., 2006)	2006	136: 5–9 years old	Primary school students	2 weeks	Kappa statistic; Pearson's correlation coefficient	Kappa 0.70. $r = 0.83$
	Allison (Allison et al., 2007)	2007	73: 5–9 years old	Primary school students	2 months	Kappa statistic; Pearson's correlation coefficient	Kappa 0.41; $r = 0.67$

LFA: low functioning autism; HFA: high functioning autism; ADHD: attention deficit hyperactivity disorder. N/A: not available.

completion of the CAST-2 was not required to be by the same parent/caregiver that completed the CAST-1. Only children with the same informant at CAST-1 and CAST-2 were included in the analyses.

2.2. Measures

The CAST was translated from English to Mandarin Chinese by the first author, a native Chinese speaker. It was back-translated by two Chinese–English bilingual speakers, not involved with autism research. The validated Taiwanese version of the CAST was used as a reference for language adjustment. In order to be culturally appropriate, the language adjustments were conducted first through discussion within a group of experts in ASC in Beijing. The Mandarin CAST was initially piloted with ten Chinese parents whose children were between 5 and 10 years of age, and selected from outpatients in the Paediatric Department of Peking University First Hospital (PUFH). The final version was back-translated and sent to the UK authors to approve. The original CAST is a 37-item parental questionnaire, of which 31 items are scored. For each scored item, one point is assigned for an ASC-positive response and 0 for an ASC-negative response. Thus, the CAST total score ranges from 0 to 31 (Baron-Cohen et al., 2009). Scoring of the Mandarin CAST was identical to the original CAST. Some items are reverse scored so that not all 'yes' responses score 1.

2.3. Procedure

Ethical approval for this research was sought from the Ethics Committee in PUFH and as well as the Cambridge University Psychology Research Ethics Committee. Consent forms were distributed with the CAST-1 to the parents of children in two schools. It contained an explanation of the purpose and procedure of the study as well as reassurances about confidentiality. Only the parents who provided consent were contacted further for this study.

2.4. Data analysis

All analyses were conducted using STATA 10.0. For each individual, the maximum score was calculated by recoding missing items to one (ASC-positive score). The minimum score was calculated by recoding missing items to zero (ASC-negative score). Initial analyses were undertaken using the minimum score. Agreement between scores on the CAST-1 and CAST-2 was assessed by treating the data in three ways:

1. in two score categories (<15 versus ≥ 15),
2. in three score categories (≤ 11, 12–14, ≥ 15) and
3. as a whole scale.

The main outcome for test–retest reliability was a measure of agreement. Cohen's kappa investigates the extent to which there is agreement other than that expected by chance expressed as a ratio to the maximum possible agreement (Cohen, 1968). Cohen's kappa = $(P_o - P_e)/(1 - P_e)$, where P_o is the observed agreement and P_e is the expected agreement which is calculated by multiplying the row total by the column total divided by the overall total (Cohen, 1968).

Overall agreement was calculated into a binary categorisation (<15 versus ≥ 15) as: $P_o = (a + d)/N$ (letters refer to Table 3) and $P_e = ((a + b)/N^*(a + c)/N + (c + d)/N^*(b + d)/N))/N$. Agreement was calculated for scoring positive for ASC (≥ 15): $P_{s+} = 2d/(2d + b + c)$, as well as both negative for ASC: $P_{s-} = 2a/(2a + b + c)$. This is the conditional probability, given that one of the scores was ≥ 15 or <15, the other would be as well (Allison et al., 2007). Exact binomial confidence intervals were calculated for these proportions. Marginal heterogeneity was assessed using an exact binomial test. The null hypothesis of the exact binomial test was that the marginal proportions were equal, which indicated that the children had the same marginal probability to move down a score group as well as up a score group over time. This is to test whether the proportion of b out of $b + c$ or the proportion of c out of $b + c$ equals to 0.5. Since a two-sided exact binomial test was applied, the probability was doubled.

The next analyses were conducted to evaluate the reliability of the Mandarin CAST using three score groups (≤ 11, 12–14 and ≥ 15). Both the kappa coefficient and weighted kappa coefficient were calculated. The latter took into account that movement across two score groups as a result of the change in the CAST-2 was more important than movement across one score group. Standard weights for agreement were applied using linear weights: 1 for no change of score group, 0.5 for change of one group, and 0 for change of two score groups (Cohen, 1968). According to the standard interpretation of Cohen's kappa, the reliability between 0.60 and 0.80 is considered to be good reliability (Altman, 1991).

Because the cut-offs for the sampling of the Mandarin CAST are still provisional, it was sensible to analyse the reliability of the Mandarin CAST as a whole scale. The Mandarin CAST score was treated as a continuous variable. Descriptive statistics were provided on the score distribution at the CAST-1 and the CAST-2. Since the distribution of scores did not follow a normal distribution, non-parametric statistical tests were used for analyses. The association between scores on the CAST-1 and CAST-2 was examined by calculating a Spearman's rank correlation coefficient. The difference between the scores on the CAST-1 and CAST-2 was also examined by the Wilcoxon signed rank test to verify the association. This approach was adopted because the correlation coefficients and their significance can only justify that the two measures are related but do not necessarily agree with each other. Therefore, there may be perfect correlation but no agreement (Bland & Altman, 1986). Thus, the correlation coefficients provide limited information because two measures can be perfectly correlated but biased

with respect to one another (Allison, 2009). In this study, because the CAST-1 and the CAST-2 were used to measure the same autistic features, they would be expected to be highly related. Thus, both Spearman's rank and Wilcoxon signed rank tests were adopted.

The difference between scores was plotted against the mean score, together with limits of agreement (Bland & Altman, 1986). The differences between scores of the two tests were calculated with their mean and standard deviation. The limits of agreements are the mean difference between the test scores plus or minus 1.96 standard deviations.

Three sensitivity analyses were carried out to investigate the effect of missing data:

1. All missing data at CAST-1 and CAST-2 were recoded as one to give a maximum score.
2. A mid-point score for each individual was generated, which was the average of the maximum and minimum score (rounded up to the nearest whole number). The analyses were conducted using the mid-point score.
3. The analyses were repeated using the minimum score of CAST-1 and the maximum score of CAST-2. This approach was to investigate the most extreme effect of missing data on the observed difference in scores.

3. Results

3.1. The study sample

In total, the parents or caregivers of 103 children completed both CASTs. $N = 70$ children with two CASTs completed by the same informant took part. Parents of $N = 59$ children (84.3%) completed both CASTs with no missing data, nine questionnaires (12.9%) had one item missing, and another two questionnaires had three or four items missing. The median age of the 70 children was 8.4 years old (range: 6.3–11.2). The mean IQ was 114 (range: 84–143). There were 36 boys and 34 girls (male:female = 1.06:1). Forty-four children (64%) were born in Beijing while the others were born in other regions in mainland China. Fifty-three children (76%) were an only child while 13 children (19%) had one brother or sister.

3.2. Two score categories

Agreement between the minimum score on the CAST-1 and CAST-2 was examined first by categorising children into two score groups (<15 and ≥15) (Table 2). The kappa statistic for binary categorisation showed that there was good agreement between the scores (kappa = 0.64, $p < 0.001$) when applying Landis's categorisation (Landis & Koch, 1977). The overall agreement of categorising an individual in the high score group (≥15) in both tests was 88.6% (95% CI: 79, 95). The specific agreement P_{s+} in the ≥15 category was 71% (95% CI: 59, 82). The specific agreement P_{s-} in the <15 score group was 93% (95% CI: 84, 98). Marginal heterogeneity was indicated ($X \sim B_{in}$ (8, 3)) two-sided, $p = 0.73$. This suggested that the differences in marginal proportions were not significant, so children were no more likely to move down a score group than they were to move up a score group.

3.3. Three score categories

Examining all score groups separately, 44 children (63%) did not move score groups, while 22 children (31%) moved down a score group. Among these, four children (6%) moved from the ≥15 to the 12–14 score group and 18 (26%) children moved from 12–14 to the ≤11 score group. Three children (4%) moved up a score group and all of them moved from the ≤11 to 12–14 score group. One child (1%) moved two score groups, from the ≥15 to the ≤11 score group (see Table 3).

Table 2
Agreement between CAST-1 and CAST-2 (<15 versus ≥15).

		CAST-2		
		<15	≥15	Total
CAST-1	<15	52 (a)	3 (b)	55 ($a+b$)
	≥15	5 (c)	10 (d)	15 ($c+d$)
	Total	57 ($a+c$)	13 ($b+d$)	70 (N)

Table 3
Agreement between the CAST-1 and CAST-2 (≤11, 12–14, ≥15).

		CAST-2			
		≤11	12–14	≥15	Total
CAST-1	≤11	17	0	0	17
	12–14	18	17	3	38
	≥15	1	4	10	15
	Total	36	21	13	70

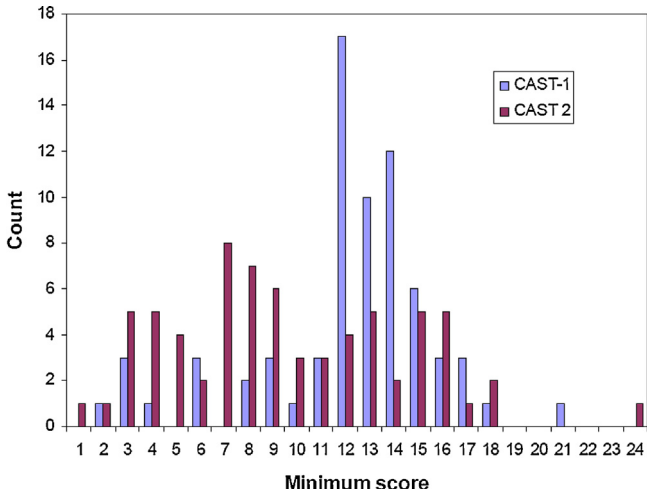

Fig. 1. Distribution of minimum scores on CAST-1 and CAST-2.

The overall agreement in the categorisation among three score groups (≤11, 12–14, ≥15) was 62.9% (95% CI: 50%, 74%). The weighted kappa showed there was moderate test–retest reliability (kappa = 0.53, $p < 0.001$).

3.4. Whole scale

The median score on the CAST-1 was 13 (range: 2–21). The median score on the CAST-2 was 11 (range: 2–24) (see Fig. 1). The Spearman's correlation coefficient between the two scores was 0.73 ($p < 0.001$). The Wilcoxon signed rank test treated the scores on the CAST-1 and CAST-2 in each individual as a test pair. The test hypothesis was that the difference between the CAST-1 and CAST-2 in each individual was equal to zero. The test statistic showed there was a significant difference between test pairs ($p = 0.0002$). The median difference between two test scores was −0.5 (IQR: −4, 0; range: −7, 3). More children (50.0%) scored lower at the time of the CAST-2 than at the CAST-1.

3.5. Change in item endorsement between two CASTs

The frequency of change including the change direction in each scorable item is shown in Table 4. All items changed in endorsement between the two CASTs. Of the 31 items, item endorsement changed in 3 items in 10% of the sample, 6 items changed in 11–20% of the sample, 8 items changed in 21–30% of the sample, 10 items changed in 31–40%, and 4 items changed in more than 40% of the sample.

3.6. Sensitivity analysis

Using different scores, three sensitivity analysis were conducted. In all three sensitivity analyses, the kappa statistic for agreement in two groups was the same with using minimum score (kappa = 0.64, $p < 0.001$) and the overall agreement was 88.6% (95% CI: 79, 95). Children were no more likely to move down a score group than to move up a score group. The overall agreement in the categorisations among three score groups (≤11, 12–14, ≥15) ranged from 58.6% to 62.9%. The weighted kappa showed the test–retest reliability was moderate (kappa = 0.48–0.53, $p < 0.001$). The median score on the CAST-1 was 13 (range: 2–21). The median score on the CAST-2 was 11 (range: 2–24). The Spearman's correlation coefficient between the two scores ranged from 0.70 to 0.73. The Wilcoxon signed rank test showed there was a significant difference between test pairs ($p < 0.001$).

4. Discussion

4.1. Main findings

This study is the first to investigate the reliability of the Mandarin CAST as a screening instrument for ASC in the Chinese population. Within an average of three months between completion of the two CAST questionnaires, the test–retest reliability across two score groups (<15 versus ≥15) was good (kappa = 0.64, $p < 0.001$). The test–retest reliability across three categories was moderate (weighted kappa = 0.53, $p < 0.001$). In addition, when treated as a whole scale, this study found a significant correlation between the CAST-1 and CAST-2 total scores (Spearman rho = 0.73). The Wilcoxon tests found that there were possible differences in score distributions of the CAST-1 and the CAST-2.

Table 4
Frequency of change in scorable items on the Mandarin CAST.

No.	Question	No. of changes	(%)	No. of scored lower in CAST-1
1	Does s/he join in playing games with other children easily?	4	(5.7)	3
5	Is it important to him/her to fit in with the peer group?	7	(10.0)	5
16	Does s/he often bring you things s/he is interested in to show you?	7	(10.0)	3
10	Does s/he find it easy to interact with other children?	9	(12.9)	3
13	Does s/he mostly have the same interests as his/her peers?	12	(17.1)	10
17	Does s/he enjoy joking around?	12	(17.1)	8
2	Does s/he come up to you spontaneously for a chat?	13	(18.6)	10
11	Can s/he keep a two-way conversation going?	14	(20.0)	11
27	Does s/he make normal eye-contact?	14	(20.0)	11
21	Are people important to him/her?	15	(21.4)	7
6	Does s/he appear to notice unusual details that others miss?	16	(22.9)	7
24	Does s/he play imaginatively with other children, and engage in role-play?	16	(22.9)	10
19	Does s/he appear to have an unusual memory for details?	17	(24.3)	6
30	Does s/he sometimes say "you" or "s/he" when s/he means "I"?	17	(24.3)	16
14	Does s/he have an interest which takes up so much time that s/he does little else?	18	(25.7)	9
34	Does s/he try to impose routines on him/herself, or on others, in such a way that it causes problems?	19	(27.1)	16
36	Does s/he often turn conversations to his/her favourite subject rather than following what the other person wants to talk about?	20	(28.6)	14
28	Does s/he have any unusual and repetitive movements?	22	(31.4)	17
35	Does s/he care how s/he is perceived by the rest of the group?	22	(31.4)	12
20	Is his/her voice unusual (e.g., overly adult, flat, or very monotonous)?	23	(32.9)	16
37	Does s/he have odd or unusual phrases?	23	(32.9)	19
8	When s/he was 3 years old, did s/he spend a lot of time pretending (e.g., play acting being a superhero, or holding teddy's tea parties)?	25	(35.7)	12
25	Does s/he often do or say things that are tactless or socially inappropriate?	25	(35.7)	17
15	Does s/he have friends, rather than just acquaintances?	26	(37.1)	17
31	Does s/he prefer imaginative activities such as play-acting or story-telling, rather than numbers or lists of facts?	27	(38.6)	12
7	Does s/he tend to take things literally?	28	(40.0)	21
29	Is his/her social behaviour very one-sided and always on his/her own terms?	28	(40.0)	20
9	Does s/he like to do things over and over again, in the same way all the time?	29	(41.4)	20
18	Does s/he have difficulty understanding the rules for polite behaviour?	29	(41.4)	20
32	Does s/he sometimes lose the listener because of not explaining what s/he is talking about?	31	(44.3)	23
23	Is s/he good at turn-taking in conversation?	33	(47.1)	26

4.2. Limitations

One limitation of this study was that although the high participation rate (97%) should have ensured the representativeness of the study sample for the population in those two schools. However, Beijing may not be nationally representative of the Chinese population due to its special political and economic status (National Bureau of Statistics of China, 2012), and therefore these results may not be generalisable to the Chinese population as a whole. Response bias could have been introduced: parents who agreed to participate in further assessment may be those who were more concerned about their children's social and communication ability than the non-responders. The time gap between the CASTs was between two to four months and was not known precisely for each child. This is a short period but no major developmental changes would be expected to occur within this timeframe. The two CASTs were conducted in different settings; the CAST-1 was taken by students back home for their parents to complete while the CAST-2 was completed during the diagnostic assessment phase in a hospital.

4.3. Comparison between the Mandarin CAST and other instruments

Using similar research methodology, the test–retest reliability of the Mandarin CAST reported in this study was higher than the UK study in a high scoring sample (Allison et al., 2007). The differences in those two study samples should be acknowledged. Although both studies were conducted in a high scoring sample, the UK study did not include children who scored $\leq$11 while this study included children from all three score groups.

The reported test–retest reliability of some screening instruments for ASC was higher than the Mandarin CAST (see Table 1). However, the current study adopted a different sampling strategy from previous studies of other instruments. This study was conducted in a high scoring sample within which 78.6% of the children scored above or around the cut-off. The difference in research methodology as well as the analytical methods between this study and previous studies made it difficult to compare their results directly.

4.4. Possible cultural influences

Despite the differences in study samples, this study found similar results to the UK study conducted in a high scoring sample. Both studies found that children were no more likely to move down a score group than move up in the CAST-2 compared with the CAST-1. It is possible that the child's behaviours rated as ASC-positive by the parents at CAST-1 were no longer noticed or reported by parents when they completed CAST-2. It is also possible some parents may have learned about the study was to examine a tool for autism from other parents whose children had already completed the assessment. It has been suggested that children with other psychiatric conditions in China may experience stigma from society (Ling, Mak, & Cheng, 2010; Mak & Kwok, 2010). Since in the Chinese culture, the academic achievement of a child is considered as the most important expectation by parents, it has been reported that children with an autism diagnosis are generally excluded by mainstream schools in mainland China. Thus, it is possible that even the parents had concerns about their children's social and communication skills at CAST-1, they might have tried to complete the CAST-2 differently in order to demonstrate that their children did not have any problems. This could have contributed the lower score on the CAST-2. In addition, adapting screening instruments developed in one culture for another culture is not without difficulties. This is because the recognition of autistic traits in the original culture may have a specific set of behavioural norms and expectations, which are not necessarily the same as the culture in the adopted country (Wallis & Pinto-Martin, 2008). For example, one of the features for case identification of autism is the eye contact. However, looking directly into someone's eye is not as common in Asian culture as it is considered shameful (Le Roux, 2002).

4.5. Conclusion and future directions

The test–retest reliability of the Mandarin CAST is moderate to good in a Chinese population. These data provide some evidence for the Mandarin CAST to be recommended as a candidate screening instrument for ASC in epidemiological studies in mainland China. Cultural aspects may be important in the adoption of a Western screening instrument for a Chinese population. Future research should use the Mandarin-CAST in a larger general population with the same informant, to further investigate the test–retest reliability.

Acknowledgements

We are grateful to all families for their participation in this study. We are indebted to Zhang Zhixiang for his advice and support throughout this study and to Yu-tzu Wu and Stephen Sharp for the discussions on data analysis. XS was supported by the Waterloo Foundation, Cambridge Commonwealth Trust and Clare Hall during the period of this study, and CA, SBC, and BA were supported by the MRC UK. This study was conducted in association with the NIHR CLAHRC for Cambridgeshire and Peterborough NHS Foundation Trust. FM was funded by the MRC UK.

References

Allison, C. (2009). *The quantitative checklist for autism in toddlers (Q-CHAT)*. Cambridge, UK: University of Cambridge (PhD).
Allison, C., Williams, J., Scott, F., Stott, C., Bolton, P., & Baron-Cohen, S. (2007). The childhood Asperger syndrome test (CAST): Test–retest reliability in a high scoring sample. *Autism, 11*(2), 177–190.
Altman, D. (1991). *Practical statistics for medical research* (1st ed.). London: Chapman & Hall.
American Psychiatric Association. (2000). Diagnostic and statistical manual of mental disorders. *DSM-IV-TR*, Washington DC: American Psychiatric Association.
Baron-Cohen, S., Scott, F. J., Allison, C., Williams, J., Bolton, P., & Matthews, F. E. (2009). Prevalence of autism-spectrum conditions: UK school-based population study. *The British Journal of Psychiatry, 194*(6), 500–509.
Baron-Cohen, S., Wheelwright, S., Cox, A., Baird, G., Charman, T., & Swettenham, J. (2000). Early identification of autism by the checklist for autism in toddlers (CHAT). *Journal of Royal Society of Medicine, 93*(10), 521–525.
Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet, 1*(8476), 307–310.
Bolte, S., Crecelius, K., & Poustka, F. (2000). The questionnaire on behaviour and social communication (VSK): An autism screening instrument for research and practice. *Diagnostica, 46*(3), 149–155.
Bolte, S., Poustka, F., & Constantino, J. N. (2008). Assessing autistic traits: Cross-cultural validation of the social responsiveness scale (SRS). *Autism Research, 1*(6), 354–363.
Bryson, S. E., Rogers, S. J., & Fombonne, E. (2003). Autism spectrum disorders: Early detection, intervention, education, and psychopharmacological management. *Canadian Journal of Psychiatry, 48*(8), 506–516.
Chlebowski, C., Green, J. A., Barton, M. L., & Fein, D. (2010). Using the childhood autism rating scale to diagnose autism spectrum disorders. *Journal of Autism and Developmental Disorders, 40*(7), 787–799.
Cohen, J. (1960). A coefficient of agreement for nominal scales. *Education and Psychological Measurement, 20*, 37–46.
Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70*(4), 213–220.
Dover, C. J., & Le, C. A. (2007). How to diagnose autism. *Archives of Diseases in Childhood, 92*(6), 540–545.
Ehlers, S., & Gillberg, C. (1993). The epidemiology of Asperger syndrome. A total population study. *Journal of Child Psychology and Psychiatry, 34*(8), 1327–1350.
Ehlers, S., Gillberg, C., & Wing, L. (1999). A screening questionnaire for Asperger syndrome and other high-functioning autism spectrum disorders in school age children. *Journal of Autism Developmental Disorders., 29*(2), 129–141.
Fombonne, E. (2009). Epidemiology of pervasive developmental disorders. *Pediatrics in Review, 65*(6), 591–598.
Gau, S. S. F., Lee, C. M., Lai, M. C., Chiu, Y. N., Huang, Y. F., & Kao, J. D. (2011). Psychometric properties of the Chinese version of the social communication questionnaire. *Research in Autism Spectrum Disorders, 5*(2), 809–818.
Gilliam, J. E. (2003). *Gilliam Asperger's disorder scale: Examiner's manual*. Austin: ProEd.
Goodman, R., & Minne, C. (1995). Questionnaire screening for comorbid pervasive developmental disorders in congenitally blind children: A pilot study. *Journal of Autism and Developmental Disorders, 25*(2), 195–203.
Kamio, Y. (2007). Early detection of and diagnostic tools for Asperger's disorder. *Nippon Rinsho/Japanese Journal of ClinIcal Medicine, 65*(3), 477–480.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.
Le Roux, J. (2002). Effective educators are culturally competent communicators. *Intercultural Education, 13*(1), 37–48.
Ling, C. Y. M., Mak, W. W. S., & Cheng, J. N. S. (2010). Attribution model of stigma towards children with autism in Hong Kong. *Journal of Applied Research in Intellectual Disabilities, 23*(3), 237–249.
Lord, C., Rutter, M., DiLavore, P., & Risi, S. (2001). *Autism diagnostic observation schedule (ADOS)*. Los Angeles, CA: Western Psychological Services.
Mak, W. W., & Kwok, Y. T. (2010). Internalization of stigma for parents of children with autism spectrum disorder in Hong Kong. *Social Science and Medicine, 70*(12), 2045–2051.
National Bureau of Statistics of China, 2011/2012. The national statistics on population. From http://www.stats.gov.cn/.
Pereira, A., Riesgo, R. S., & Wagner, M. B. (2008). Childhood autism: Translation and validation of the childhood autism rating scale for use in Brazil. *Journal of Pediatrics (Rio J), 84*(6), 487–494.
Pine, E., Luby, J., Abbacchi, A., & Constantino, J. N. (2006). Quantitative assessment of autistic symptomatology in preschoolers. *Autism, 10*(4), 344–352.
Rutter, M., LeCouteur, A., & Lord, C. (2003). *Autism diagnostic interview-revised manual*. Los Angeles, CA: Western Psychological Services.
Scott, F. J., Baron-Cohen, S., Bolton, P., & Brayne, C. (2002). The CAST (childhood Asperger syndrome test): Preliminary development of a UK screen for mainstream primary-school-age children. *Autism, 6*(1), 9–31.
Sun, X., & Allison, C. (2009). A review of the prevalence of autism spectrum disorder in Asia. *Research in Autism Spectrum Disorder, 4*(2), 156–167.
Sun, X., Allison, C., Matthews, F., Zhang, Z., Auyeung, B., & Baron-Cohen, S. (2012). *An exploration of the underdiagnosis of autism in mainland China using screening and diagnostic instruments*. Cambridge: The Cambridge Autism Research Centre (submitted for publication).
Tang, Y., Guo, Y., Rice, C., Wang, E., & Cubells, Y. J. F. (2010). Introduction of the "gold standard" diagnostic instrument (autism diagnostic observation scale). *International Journal of Psychiatry, 37*(1), 38–40.
Vostanis, P., Smith, B., Chung, M. C., & Corbett, J. (1994). Early detection of childhood autism: A review of screening instruments and rating scales. *Child: Care, Health and Development, 20*(3), 165–177.
Wallis, K. E., & Pinto-Martin, J. (2008). The challenge of screening for autism spectrum disorder in a culturally diverse society. *Acta Paediatrica, 97*(5), 539–540.
Wang, Y., Wang, G., & Wang, Y. (2003). Analysis of childhood autism by using Clancy autism behavior scale and autism behavior checklist. *Journal of Shandong University (Health Science), 41*(2), 213–214.
Williams, J. (2003). *Screening for autism spectrum disorders*. University of Cambridge.
Williams, J., Allison, C., Scott, F., Stott, C., Bolton, P., & Baron-Cohen, S. (2006). The childhood Asperger syndrome test (CAST): Test–retest reliability. *Autism, 10*(4), 415–427.
Williams, J., Scott, F., Stott, C., Allison, C., Bolton, P., & Baron-Cohen, S. (2005). The CAST (childhood Asperger syndrome test): Test accuracy. *Autism, 9*(1), 45–68.
World Health Organisation. (1993). *The ICD-10 Classification of Mental and Behavioural Disorder: Diagnosis Criteria for Research*. Geneva: World Health Organisation.
Yang, X. L., Huang, L. Q., Jia, M. X., & Chen, S. (1993). Validation study of autism behavior checklist. *Chinese Journal of Mental Health, 7*(6), 279–280.
Yen, M., & Lo, L. H. (2002). Examining test–retest reliability: An intra-class correlation approach. *Nursing Research, 51*(1), 59–62.
Zhang, X., & Ji, C. Y. (2005). Autism and mental retardation of young children in China. *Biomedical and Environmental Sciences, 18*(5), 334–340.