autism © 2006
SAGE Publications
and The National
Autistic Society
Vol 10(4) 415–427; 066612
1362-3613(200607)10:4

The Childhood Asperger Syndrome Test (CAST)

Test–retest reliability

JO WILLIAMS *University of Cambridge, UK*
CARRIE ALLISON *University of Cambridge, UK*
FIONA SCOTT *University of Cambridge, UK*
CAROL STOTT *University of Cambridge, UK*
PATRICK BOLTON *University of Cambridge, UK*
SIMON BARON-COHEN *University of Cambridge, UK*
CAROL BRAYNE *University of Cambridge, UK*

ABSTRACT The Childhood Asperger Syndrome Test (CAST) is a 37-item parental self-completion questionnaire to screen for autism spectrum conditions in research. Good test accuracy was demonstrated in studies with primary school aged children in mainstream schools. The aim of this study was to investigate the test–retest reliability of the CAST. Parents of 1000 children in years 1–6 in five mainstream primary schools in Cambridgeshire received the CAST. A second identical questionnaire was posted to respondents after approximately 2 weeks. Both mailings generated 136 responses. Agreement above and below a screening cut-point of 15 was investigated. The kappa statistic for agreement (< 15 versus ≥ 15) was 0.70, and 97 percent (95 percent CI: 93–99 percent) of children did not move across the cut-point of 15. The correlation between the two test scores was 0.83 (Spearman's rho). The CAST has shown good test–retest reliability, and now requires further investigation in a high-scoring sample.

ADDRESS *Correspondence should be addressed to:* JO WILLIAMS, *Department of Public Health and Primary Care, University of Cambridge, Forvie Site, Robinson Way, Cambridge CB2 2SR, UK. e-mail: j.g.williams.97@cantab.net*

KEYWORDS *Asperger syndrome; autism; childhood screening; pervasive developmental disorder; reliability*

Introduction

The Childhood Asperger Syndrome Test (CAST)[1] is a 37-item parental self-completion questionnaire designed to screen for autism spectrum conditions. It can be used to identify possible cases of autism spectrum

conditions in the general population in research. The CAST is particularly designed to identify Asperger syndrome and the subtler manifestations of autism spectrum conditions amongst primary school aged children attending mainstream schools. In previous population-based studies, the CAST has shown good sensitivity and specificity when using a score cut-point of ≥ 15 (maximum possible score 31) (Scott et al., 2002; Williams et al., 2005). The positive predictive value was moderate, and the response rate was 26 percent in the general population (Williams et al., 2005). A test that has been shown to have good validity must also be shown to be reliable. It is important to establish the reliability of a test so that those who use the test are aware of the extent to which measurement error is present and aware of its repeatability and stability over time.

Test–retest reliability has been investigated for other autism screening tests. Of the numerous published autism screening tests, four tests have reported test–retest reliability (Table 1). The data in these four studies were treated in different ways, which hinders comparison between the tests. The

Table 1 **Investigations of test–retest reliability in autism screening tests**

Screen and reference	*Sample size and population*	*Time between tests*	*Results*
CSBQ (Luteijn et al., 2000)	21 from a mixed sample	*c.* 1 month	Intraclass correlation, whole scale = 0.90 Individual scales ranged from 0.32 to 0.85
AQ (Baron-Cohen et al., 2001)	17 from general population sample	2 weeks	Pearson $r = 0.7$ ($p = 0.0002$) Paired *t*-test for difference between scores, $t(16) = 0.3$, $p = 0.75$
ASSQ (Ehlers et al., 1999)	65 from 6–17 year-olds with ASD, ADHD or learning disabilities at a child neuro-psychiatric clinic	2 weeks	Teacher version, $r = 0.94$ ($p < 0.0001$) Parent version, $r = 0.96$ ($p < 0.0001$). No significant difference found using paired *t*-test
ASDASQ (Nylander and Gillberg, 2001)	38 from adult psychiatric outpatients	11–13 months	Kendell's tau = 0.69 ($p < 0.001$) Spearman's rho = 0.82 ($p < 0.001$) Agreement across items ranged from 74% to 87%, and kappa from 0.22 to 0.65

Children's Social Behaviour Questionnaire (CSBQ: Luteijn et al., 2000) was seen as multiple categories and tested using repeated-measures ANOVA, giving rise to an intra-class correlation. This is equivalent to a kappa statistic when there are two categories, and to a weighted kappa when there are more than two categories (Norman and Streiner, 2000, p. 221). The Autism Spectrum Quotient (AQ: Baron-Cohen et al., 2001) and the Autism Spectrum Screening Questionnaire (ASSQ: Ehlers et al., 1999) were analysed using a Pearson's correlation coefficient and a paired *t*-test to test for a difference between scores. The Autism Spectrum Disorder in Adults Screening Questionnaire (ASDASQ: Nylander and Gillberg, 2001) was analysed using non-parametric methods and the association between the measures was investigated using Spearman's rho and Kendall's tau. The items were treated as categorical data and overall agreement and kappa statistics were calculated for each item. All these four tests show good test–retest reliability, although the time interval for the ASDASQ was very long at 11–13 months (Table 1).

The aim of this study was to investigate the test–retest reliability of the CAST in a population sample. To enable comparison with other screening tests, the data on the CAST were analysed both according to the score categories currently used for sampling in research and by treating the CAST as a whole scale.

Methods

Data collection

A letter was sent to five mainstream primary schools in Cambridgeshire to ask if they would take part in the test–retest reliability study. This was followed up by telephone. As only three of the five schools approached agreed to participate, a further two schools were recruited.

Parents of 1000 children in years 1–6 (age 5–11 years) in the five schools received the CAST. The questionnaires (TEST-1) were distributed in class registration, and returned to the research team by parents using a Freepost envelope. Minimum personal information was collected and it was verified that the same person had completed both questionnaires. Respondents to TEST-1 were sent a second questionnaire direct to their home approximately 2 weeks after the first; if parents did not respond within this time, a second questionnaire was sent approximately 1 week after receiving the first questionnaire. The second mailing (TEST-2) was identical to the first with the exception of the covering letter. Data were also collected on the exact time interval between the return of the two questionnaires.

Analysis

The questionnaires were scored and missing data assessed. Initial analyses were undertaken comparing the TEST-1 and TEST-2, treating all missing data as if the item score would have been zero (observed score). All analyses were carried out including only individuals with two CAST questionnaires completed by the same respondent. The characteristics of responders and non-responders to TEST-2 were investigated to assess if there was response bias. Agreement between scores from TEST-1 and TEST-2 was assessed treating the data in three ways:

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

The main outcome for test–retest reliability was a measure of agreement, kappa, for the group with scores ≥ 15. Kappa investigates the agreement beyond that which would be expected by chance expressed as a ratio to the maximum possible agreement beyond chance, namely $(P_o - P_e)/(1 - P_e)$, where P_o is the observed agreement and P_e is the expected agreement (Cohen, 1968). Standard interpretations state that very good agreement is indicated by kappa > 0.8, whilst kappa > 0.6 indicates good agreement (Altman, 1991, p. 404). Where there was more than one score category, a weighted kappa was used, which accounts for movement across one score group being less important than movement across two. Standard weights were used: 1 for no change of score group, 0.5 for change of one group, and 0 for change of two groups (Cohen, 1968, cited in StataCorp, 2001).

Overall agreement of classification into a binary categorization (< 15 versus ≥ 15) was calculated as $P_o = (a + d)/N$ (letters *a*–*d* refer to Table 2). Overall agreement into three score categories was calculated as the number of children scoring in the same category at both tests as a proportion of all the children tested. Specific agreement in two score categories was calculated for scoring in the score group of interest: $P_{s+} = 2d/(2d + b + c)$. This is the conditional probability, given that one of the scores was ≥ 15, that the other would be as well. Specific agreement was also calculated for

Table 2 **Score categories for calculating agreement indices**

		TEST-2		
		< 15	*≥ 15*	*Total*
TEST-1	< 15	*a*	*b*	*a + b*
	≥ 15	*c*	*d*	*c + d*
	Total	*a + c*	*b + d*	*N*

scoring outside the score group of interest: $P_{s-} = 2a/(2a + b + c)$. Exact binomial confidence intervals were calculated on all proportions. The degree of marginal heterogeneity was assessed using an exact binomial test of each marginal proportion occurring, that is the probability of b given N of $b + c$ ($X \sim \text{Bin}(b + c, b)$). A two-sided exact binomial test was undertaken, so the probability was doubled. This tested the null hypothesis that the marginal proportions were equal, that is, that the children were as likely to move down a score group as up a group over time.

As the score boundaries for sampling are still provisional, it was valuable also to analyse the reliability of the CAST as a whole scale. Descriptive statistics of the score distribution at TEST-1 and TEST-2 were given. It was appropriate to use non-parametric statistical measures as the distribution of scores did not follow a normal distribution. A Spearman's rho correlation coefficient between TEST-1 and TEST-2 was calculated to describe the association between the scores. It should be noted that correlation coefficients provide limited information as two measures can be perfectly correlated but still biased with respect to one another, that is there may be perfect correlation but no agreement (Bland and Altman, 1986). As a result a Wilcoxon signed-rank test was used to indicate if there was a significant difference between the two scores.

A sensitivity analysis was carried out to investigate the effect of missing data. The analyses were repeated after missing items were recoded to one (maximum score). All analyses were carried out using STATA (version 7.0).

Results

Response and missing data

Two pairs of questionnaires were excluded: in one pair a questionnaire had an entire page of the questionnaire missing; and in another pair one questionnaire was returned blank. The response rate, after exclusions, to TEST-1 was 28 percent ($n = 282$) and to TEST-2 was 48 percent ($n = 136$). There were no significant differences between participants who responded or did not respond to TEST-2 in the child's age (t-test, $p = 0.5$) or gender (χ^2, $p = 0.9$), or in whether previous concerns had been expressed over the child's development by teachers or health visitors (χ^2, $p = 0.9$). However, those responding to TEST-2 had significantly lower scores at TEST-1 (median = 3; IQR 2, 6; $n = 136$) than non-responders to TEST-2 (median = 5; IQR 3, 8; $n = 146$) (Wilcoxon rank-sum test, $p = 0.02$). The time interval between the two tests on each individual had a median 22 days (IQR 15–25, range 7–64). Overall 82 percent of the questionnaires were complete, 11 percent had one item missing, and the remaining 7 percent had between two and seven missing items.

Two score categories (< 15 versus ≥ 15)

The kappa statistic (< 15 versus ≥ 15) was 0.70 ($p < 0.0001$), showing good agreement between the results on the two tests for each child. The overall agreement of categorization across the two time points was 97 percent (95 percent CI: 93–99 percent) (Table 3). The indices of specific agreement were P_{s+} = 71 percent (95 percent CI: 42–92 percent) and P_{s-} = 98 percent (95 percent CI: 96–99.6 percent). Marginal heterogeneity was not indicated ($X \sim \text{Bin}(4, 1)$ two-sided, $p = 0.63$), showing that children were no more likely to move up than down a score group.

Three score categories (≤ 11; 12–14; ≥ 15)

Using three score categories (Table 4), it was seen that 129 children (95 percent) did not move score group. No children increased their score by two groups; three (2 percent) children increased their score by one score group; four (3 percent) children moved down one score group; and no children moved down two score groups. The overall agreement across all score categories was 95 percent (95 percent CI: 90–98 percent). The weighted kappa was 0.82 ($p < 0.0001$).

Whole scale

The score distributions for TEST-1 (median 3, IQR 2–6) and TEST-2 (median 3, IQR 2–5.5) were very similar (Figure 1), and the correlation between the two tests was 0.83 (Spearman's rho) (Figure 2). The Wilcoxon

Table 3 **Agreement between TEST-1 and TEST-2 (< 15 versus ≥ 15)**

		TEST-2		
		< 15	*≥ 15*	*Total*
TEST-1	< 15	127	3	130
	≥ 15	1	5	6
	Total	128	8	136

Table 4 **Agreement between TEST-1 and TEST-2 (≤ 11; 12–14; ≥ 15)**

		TEST-2			
		≤ 11	*12–14*	*≥ 15*	*Total*
TEST-1	≤ 11	120	0	0	120
	12–14	3	4	3	10
	≥ 15	0	1	5	6
	Total	123	5	8	136

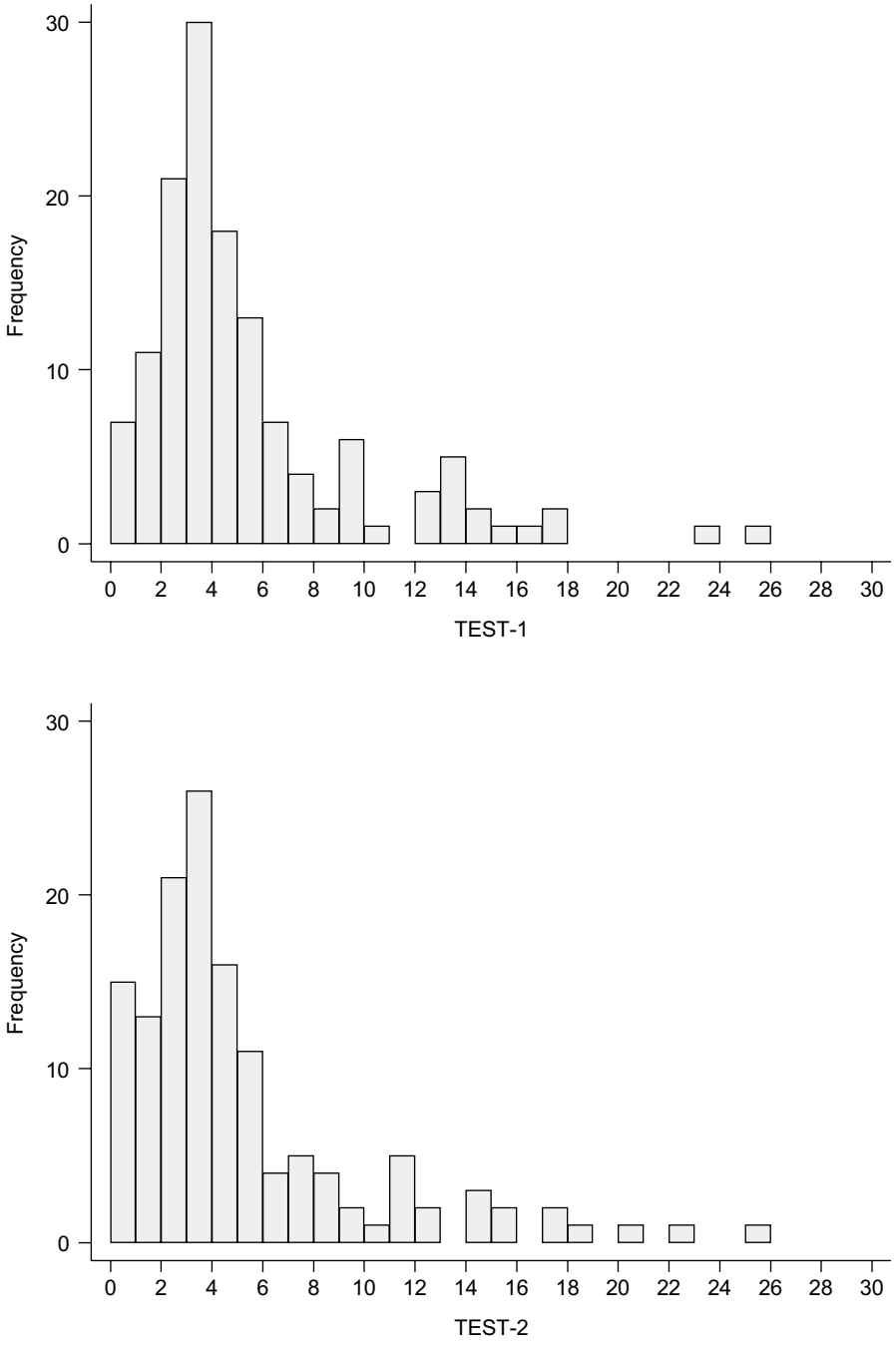

Figure 1 **Score distributions for TEST-1 and TEST-2**

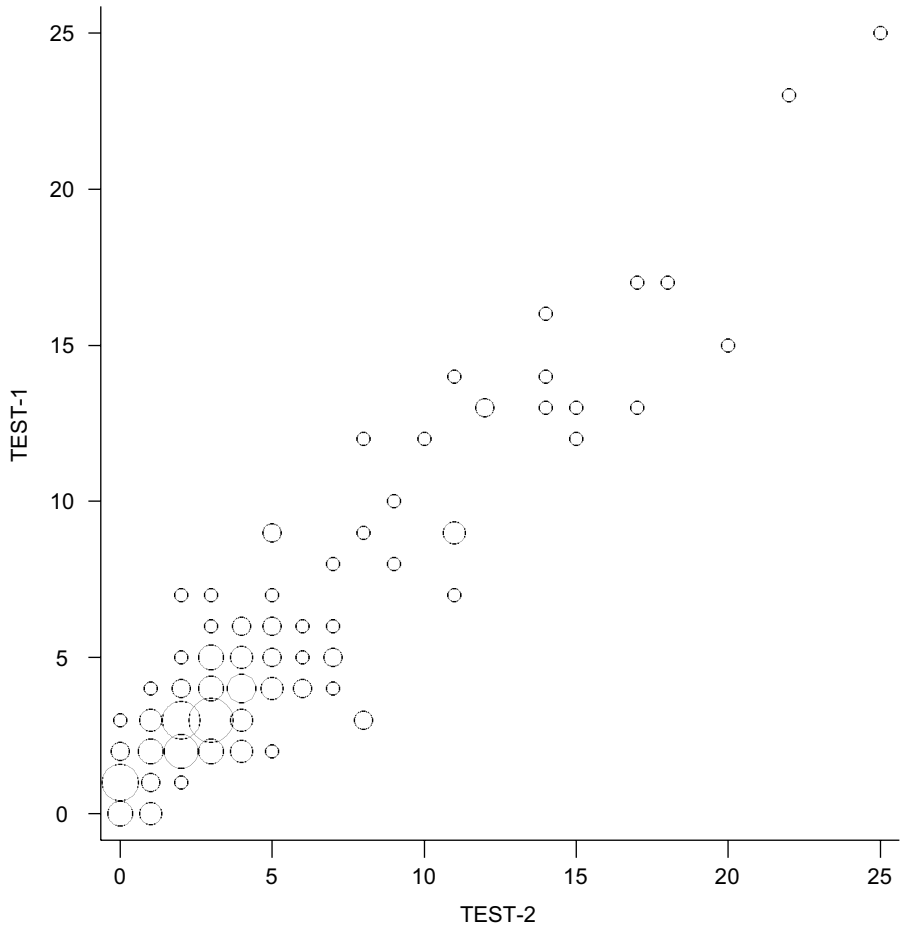

Figure 2 **Scatter plot of TEST-1 and TEST-2 (sizes of circles are proportional to the number of individuals represented)**

signed-rank test showed a possible difference between score distributions at TEST-1 and TEST-2 ($p = 0.04$). A higher proportion had a decrease in score (45.5 percent) than an increase (27.1 percent); however the majority of these changes were very small (Table 5). The difference between the pairs of test scores had a median of 0 (IQR −1 to 1, range −5 to 5).

Sensitivity analyses

When missing items were recoded as 1, the indices of agreement dropped only slightly. The kappa statistic for agreement across the < 15 and ≥ 15 score groups was 0.60 ($p < 0.0001$), and the overall agreement was 96

Table 5 **Score difference (TEST-1 minus TEST-2)**

Score difference	*Frequency*	*Percentage*	*Cumulative percentage*
–5	1	0.7	0.7
–4	4	2.9	3.7
–3	5	3.7	7.4
–2	16	11.8	19.1
–1	36	26.5	45.6
0	36	26.5	72.1
1	19	14.0	86.0
2	11	8.1	94.1
3	3	2.2	96.3
4	2	1.5	97.8
5	3	2.2	100.0

percent (95 percent CI: 91–98 percent). There was still no indication of marginal heterogeneity. When scoring in three categories the weighted kappa dropped to 0.74 ($p < 0.0001$), indicating that there was still good test–retest reliability. When using maximum score, for TEST-1 the median score was 4 (IQR 2–6.5, range 0–25), and for TEST-2 the median score was 3 (IQR 2–6, range 0–26), which were only slightly higher scores than when using the observed score. The correlation between the two test results was 0.81 (Spearman's rho), and the Wilcoxon signed-rank test now showed that the two test results were not significantly different from one another ($p = 0.20$). The median difference between scores at TEST-1 and TEST-2 was 0 (IQR −1 to 1, range −5 to 9).

Discussion

Main findings

In this population sample, there was good test–retest reliability between screen results using the CAST (TEST-1) and a CAST retest (TEST-2) within an average of 3 weeks. This was demonstrated by a kappa statistic of 0.70 (< 15 versus ≥ 15), and the finding that 97 percent of children did not move between score groups. As a cut-point of 15 continues to be used for screening in epidemiological research, it was extremely important to establish that the ≥ 15 score group had good reliability. It was also informative to find that the children were no more likely to move up or down score groups.

The 12–14 score group also has been used for sampling in epidemiological studies, to assess if any cases of autism spectrum conditions are found amongst those scoring below the cut-point of 15. It was therefore

valuable to discover that the test–retest reliability across all score groups was good, with a weighted kappa of 0.82 across three score groups.

The score cut-points on the CAST are provisional. Therefore it was important to show good correlation between the scores at TEST-1 and TEST-2 across the whole scale. Treating the CAST as a scale, there was high correlation between the test scores. The Wilcoxon signed-rank test showed a possible difference between observed scores at TEST-1 and TEST-2, but the probability was of borderline statistical significance, and may in part have reflected the finding that TEST-1 scores were lower in responders to TEST-2 than in non-responders. The difference between the score distributions reflected very small differences between the two test scores and a significant difference was not seen in the indices of agreement across score groups. When the effect of missing data was examined in the sensitivity analyses, this difference between score distributions was no longer found. It is therefore unlikely to be an important difference in terms of whether or not a child would be selected for further assessment in a study.

Whilst it is important to consider the influence of regression to the mean when using repeated measures, there is little evidence of this effect in this study. There was very little movement between score groups and across the whole scale. Indeed, very little regression to the mean was expected in this situation as the correlation between the test scores was high, and the whole population was approached rather than an extreme sample.

The CAST was found to have good test–retest reliability, comparable to that of previously published studies of autism screening tests (Table 1). For example, the correlation between tests on the ASDASQ was 0.82 (Spearman's rho) over a long time interval of 11–13 months (Nylander and Gillberg, 2001) whilst that on the CAST was 0.83. It is harder to make comparison to other tests due to the different statistical measures used, but the correlation between scores at two time points was high as in this study of the CAST. The AQ had a correlation coefficient (Pearson) of 0.7 (Baron-Cohen et al., 2001), and the ASSQ a correlation of 0.96 (parent version) for scores 2 weeks apart (Ehlers et al., 1999).

Strengths and weaknesses of the study and further research

The response to the CAST questionnaires in this study in the general population was relatively low at 28 percent and similar to that in previous studies (Scott et al., 2002; Williams et al., 2005). The characteristics of non-respondents to the CAST may be such that the reliability of the test might have been somewhat different if the whole sample had responded.

A proportion of the questionnaires had missing data. A simple method of indicating the effect of the missing data was used, which examined the

extremes of possible scores, the minimum and maximum that the individuals could have received. The agreement and reliability indices were high in both cases; therefore we can be confident that missing data did not affect the overall conclusion that the test–retest reliability of the CAST was good.

This test–retest reliability study was carried out in a large sample and is therefore likely to have generated robust results. The two questionnaires were completed by the same parent or guardian and the exact time interval between the questionnaires was known. The sample for this test–retest reliability study was from the general population in mainstream schools, and as a result only a minority of children received high scores on the CAST. The most important aspect of test–retest reliability is to establish the stability of the screen near the cutoff score of 15; therefore a test–retest exercise should be carried out in a sample enriched by individuals with high scores. This is currently being investigated in a large epidemiological study, one aim of which is to investigate the prevalence of autism spectrum conditions, and a second aim of which is to assess the performance of the CAST in identifying possible cases of autism spectrum conditions in epidemiological research. In this study, the CAST is given as a retest approximately 1–2 months after a first test to a sample of children who are receiving more detailed assessments, which consists of all the children scoring at or above 15, and a proportion of those scoring between 12 and 14 on the first CAST.

Conclusion

The CAST has shown good test–retest reliability in a large sample of children from mainstream schools. Reliability was found to be good both across the provisional score cut-points of 15 and 12, and across the scale as a whole. Further investigation of the test–retest reliability in a high-scoring sample is under way.

The CAST is not currently recommended for national or comprehensive screening in a public health or educational setting due to the lack of evidence for screening in the population (National Screening Committee Child Health Subgroup, 2001), and due to the moderate positive predictive value and the relatively low response rates (Williams et al., 2005). However, this study contributes new evidence about the reliability of the CAST, and, together with previous investigations that showed that the test had good sensitivity and specificity (Scott et al., 2002; Williams et al., 2005), suggests that the CAST is robust for screening to identify possible autism spectrum conditions in epidemiological studies.

Acknowledgements

This study was carried out collaboratively by members of the Department of Public Health and Primary Care, and the Departments of Psychiatry and Experimental Psychology, at the Autism Research Centre, University of Cambridge. We are grateful to the Shirley Foundation for their generosity in funding this study. The Shirley Foundation also funded the accompanying production of a brief television documentary ('Asperger Syndrome: A Different Mind') to help raise awareness among teachers in primary schools (www.jkp.com). Jo Williams (née Johnson) held an MRC PhD studentship. Simon Baron-Cohen was also supported by the MRC during the period of this work. We wish to thank Til Utting-Brown for assistance with data entry, and Tom Fanshawe (Centre for Applied Medical Statistics, University of Cambridge) for statistical advice. We are grateful to the schools that participated, and to the parents who gave up their time to complete the questionnaires.

Notes

1 A copy of the CAST is available in two articles (Scott et al., 2002; Williams et al., 2005) and electronically at http://www.autismresearchcentre.com/instruments/research_instruments.asp.

References

ALTMAN, D. (1991) *Practical Statistics for Medical Research*, 1st edn. London: Chapman & Hall.

BARON-COHEN, S., WHEELWRIGHT, S., SKINNER, R., MARTIN, J. & CLUBLEY, E. (2001) 'The Autism-Spectrum Quotient (AQ): Evidence from Asperger Syndrome/High-Functioning Autism, Males and Females, Scientists and Mathematicians', *Journal of Autism and Developmental Disorders* 31 (1): 5–17.

BLAND, J.M. & ALTMAN, D.G. (1986) 'Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement', *Lancet* 1 (8476): 307–10.

COHEN, J. (1968) 'Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit', *Psychological Bulletin* 70: 213–20.

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 and Developmental Disorders* 29 (2): 129–41.

LUTEIJN, E., LUTEIJN, F., JACKSON, S., VOLKMAR, F. & MINDERAA, R. (2000) 'The Children's Social Behavior Questionnaire for Milder Variants of PDD Problems: Evaluation of the Psychometric Characteristics', *Journal of Autism and Developmental Disorders* 30 (4): 317–30.

NATIONAL SCREENING COMMITTEE CHILD HEALTH SUBGROUP (2001) 'National Screening Committee Policy Position on Screening for Autism', http://www.nelh.nhs.uk/screening/child_pps/autism.html, accessed 10 January 2003.

NORMAN, G.R. & STREINER, D.L. (2000) *Biostatistics: The Bare Essentials*, 2nd edn. Hamilton, Canada: Decker.

NYLANDER, L. & GILLBERG, C. (2001) 'Screening for Autism Spectrum Disorders in Adult Psychiatric Out-Patients: A Preliminary Report', *Acta Psychiatrica Scandinavica* 103 (6): 428–34.

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.

STATACORP (2001) *Stata Statistical Software: Release 7.0*. College Station, TX: Stata Corporation.

WILLIAMS, J., SCOTT, F., STOTT, C., ALLISON, C., BOLTON, P., BARON-COHEN, S. & BRAYNE, C. (2005) 'The CAST (Childhood Asperger Syndrome Test): Test Accuracy', *Autism* 9 (1): 45–68.