autism © 2007
SAGE Publications
and The National
Autistic Society
Vol 11(2) 173–185; 075710
1362-3613(200703)11:2

The Childhood Asperger Syndrome Test (CAST)

Test–retest reliability in a high scoring sample

CARRIE ALLISON *University of Cambridge, UK*
JO WILLIAMS *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 designed to screen for high-functioning autism spectrum conditions in epidemiological research. The CAST has previously demonstrated good accuracy for use as a screening test, with high sensitivity in studies with primary school aged children in mainstream schools. This study aimed to investigate test–retest reliability of the CAST in a high scoring sample. To this end, 73 parents filled in the second CAST (CAST-2) within approximately 2 months of the first administration of the CAST (CAST-1). Agreement above and below the cut-point of 15 was investigated. The kappa statistic for agreement (<15 versus ≥15) was 0.41. It was found that 70 percent (95% CI: 58, 80) of children did not move across the cut-point of 15. The correlation between the two test scores was 0.67 (Spearman's rho). The CAST shows moderate test–retest reliability in a high scoring sample, further evidence that it is a relatively robust screening tool for epidemiological research.

ADDRESS *Correspondence should be addressed to:* CARRIE ALLISON, *Autism Research Centre, Douglas House, 18b Trumpington Road, Cambridge CB2 8AH, UK. e-mail: cla29@cam.ac.uk*

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

Introduction

The Childhood Asperger Syndrome Test (CAST)[1] is a parental questionnaire based on ICD-10 (World Health Organization, 1993) and DSM-IV (American Psychiatric Association, 1994) criteria designed to identify subtle

DOI: 10.1177/1362361307075710

manifestations of autism spectrum conditions, and in particular Asperger syndrome. The questionnaire includes 31 key items contributing to a child's total score, along with six control questions on general development. The CAST has been used in population-based epidemiological research as a screen for autism spectrum conditions in primary school aged children, and previous studies have shown reasonable sensitivity and specificity when using a screening cut-point of ≥15 (Scott et al., 2002). In this pilot study, all the children with a diagnosis of Asperger syndrome scored at 15 or above while none of the typically developing children scored above 15. The CAST has good validity (see Williams et al., 2005), and it is also important to assess its reliability and stability over time.

Test–retest reliability of the CAST was investigated in a previous study (Williams et al., 2006). The aim of that study was to examine the test–retest reliability in a population sample according to the score categories currently used for sampling in research, and also treating the CAST as a scale. The CAST showed good agreement between screen results at or above, and below, a cut-point of 15 with a kappa statistic of 0.70. This study was based on a sample from mainstream schools, and therefore only a small proportion of children scored above the cut-point on the CAST. An important aspect of test–retest reliability of a screening instrument is to establish the stability of the screen near the proposed cut-point. A comparison between test–retest reliability in published autism screening tests has been reviewed elsewhere (Williams et al., 2006). The aim of the present study is to examine the test–retest reliability of the CAST in a sample enriched by children scoring at or above the cut-point of 15, as well as a proportion of those scoring 12–14.

Methods

Data collection

A total of 11,635 CAST questionnaires were distributed via schools throughout Cambridgeshire to parents/caregivers of children aged 5 to 9 years old. The CAST was distributed in class registration and returned to the research team by parents using a Freepost envelope. Minimum personal information was collected. Those children who scored above the cut-point of 15, and those who scored between 12 and 14, were invited for a full diagnostic assessment. At the time of assessment, parents/caregivers were asked to complete a second identical copy of the CAST (CAST-2). It was intended that the CAST-2 would be completed within 2 months of the first CAST (CAST-1), and data were collected on the exact time interval between the two administrations of the CAST. Typically, the CAST-2 was given to the

parent/caregiver before the Autism Diagnostic Interview–Revised (ADI–R: Lord et al., 1994) and the Autism Diagnostic Observation Schedule (ADOS: Lord et al., 2001) assessments. Data were not collected to check that the same parent/caregiver filled in both versions of the CAST.

Analysis

The CAST-2 was scored (for a description of the scoring procedure, see http://www.autismresearchcentre.com/instruments/research_instruments.asp) and missing data were assessed. The maximum score was calculated by recoding missing items to one. The observed score treated all missing data as if the item score would have been zero. All analyses were carried out only including individuals with two CAST questionnaires. Initial analyses were undertaken using the midpoint score, that is the maximum score plus the observed score/2 (rounded up to nearest whole number) of CAST-1 and CAST-2, as this was the score used for sampling. This treats the missing data as if the individual had scored on half the items.

Agreement between scores from CAST-1 and CAST-2 was assessed by treating the data in three ways:

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

The main outcome for test–retest reliability was a measure of agreement, kappa, which investigates the agreement beyond that which would be expected by chance expressed as a ratio to the maximum possible agreement beyond chance, i.e. $(P_o - P_e)/(1 - P_e)$, where P_o is the observed agreement and P_e is the expected agreement (Cohen, 1968). Standard interpretations of kappa are as follows (Altman, 1991, p. 404):

- poor agreement: less than 0.20
- fair agreement: 0.21 to 0.40
- moderate agreement: 0.41 to 0.60
- good agreement: 0.61 to 0.80
- very good agreement: 0.81 to 1.00

Overall agreement was calculated for classification into a binary categorization (<15 versus ≥15) as $P_o = (a + d)/N$ (letters a–d refer to Table 1). Specific agreement 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 scoring outside the score group of interest: $P_{s-} = 2a/(2a + b + c)$. Exact binomial confidence intervals were calculated on these 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 applied, 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 score group across time.

Overall agreement and the kappa coefficient were calculated across all three score groups (≤11, 12–14, ≥15). In addition, a weighted kappa coefficient was calculated to take into account the fact that movement across one score group was less important than movement across two score groups. 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).

As the score boundaries for sampling are still provisional, it was valuable to analyse the reliability of the CAST as a whole scale. Descriptive statistics of the score distribution at CAST-1 and CAST-2 were given. As the distribution of scores did not follow a normal distribution, it was appropriate to use non-parametric statistical measures. A Spearman's rho correlation coefficient was calculated to describe the association between CAST-1 and CAST-2 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). In light of this, a Wilcoxon signed-rank test was used to verify whether there was a significant difference between the two test scores.

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

- All missing data at CAST-1 and CAST-2 were recoded as zero to give an observed score.
- All missing data at CAST-1 and CAST-2 were recoded as 1 to give a maximum score.
- The analyses were repeated using observed score at time 1 and maximum score at time 2, to model the most extreme effect of missing data on any observed difference in scores.

All analyses were carried out using SPSS (version 11.5) and STATA (version 7.0).

Results

A total of 74 questionnaires from the second administration of the screening test (CAST-2) were available from the assessment sample, completed

Table 1 **Agreement between CAST-1 and CAST-2 (<15 versus ≥15) (*N* = 73)**

		CAST-2		
		<15	*≥15*	*Total*
CAST-1	*<15*	20 (*a*)	4 (*b*)	24 (*a* + *b*)
	≥15	18 (*c*)	31 (*d*)	49 (*c* + *d*)
	Total	38 (*a* + *c*)	35 (*b* + *d*)	73 (*N*)

between 13 and 330 days after their original CAST (CAST-1). The median number of days between questionnaire administrations was 54 (IQR 34–110). One questionnaire was excluded on the basis that the child had been incorrectly sampled, leaving 73 questionnaires available from both administrations.

Two score categories (<15 versus ≥15)

Agreement was investigated when categorizing children into the ≥15 or <15 score group (Table 1). The kappa statistic for the binary categorization showed that there was moderate agreement between the scores at the two administrations of the CAST (kappa = 0.41, $p < 0.001$), applying Altman's (1991) categorization. The overall agreement for categorizing an individual in the ≥15 score group at both tests was 69.9 percent (95% CI: 58, 80). The specific agreement P_{s+} in the ≥15 category was 73.8 percent (95% CI: 63, 83). The specific agreement P_{s-} for categorizing the individual in the <15 score group both times was 64.5 percent (95% CI: 51, 76). Marginal heterogeneity was indicated ($X \sim \text{Bin}(22, 4)$ two-sided, $p = 0.004$), that is, children were no more likely to decrease in score group than to increase in score group, and children tended to move down a group more often than up.

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

A total of 39 (53%) individuals did not move score groups. Four children (5%) increased their CAST score so as to move up a score group (Table 2), moving from the middle to the highest score group; 19 (26%) of individuals moved down one score group; and 11 (15%) of individuals moved down two score groups. There were no children at CAST-1 in the low score group since this study was conducted on children who were sampled due to their middle or high score status at CAST-1.

The overall agreement in the categorizations across all three score groups (≤11, 12–14, and ≥15) was $P_0 = 53$ percent (95% CI: 41, 65). The weighted kappa was 0.25 ($p < 0.0001$).

Table 2 **Agreement between CAST-1 and CAST-2 (≤11, 12–14, ≥15) (*N* = 73)**

		CAST-2			
		≤11	*12–14*	*≥15*	*Total*
CAST-1	*≤11*	0	0	0	0
	12–14	12	8	4	23
	≥15	11	7	31	45
	Total	23	15	35	73

Whole scale

The scores from the CAST-1 had a median of 16 (IQR 14–19, range 12–29). The scores from the CAST-2 had a median of 14 (IQR 11–18, range 5–27) (Figures 1 and 2).

The correlation coefficient (Spearman's rho) between the two scores was 0.67, and a Wilcoxon signed-rank test showed a significant difference between test pairs (testing the null hypothesis that the difference was not equal to zero, $p < 0.001$). The median difference between scores was −2 (IQR −4–0), ranging from −10 to +5. A higher proportion had a decrease in score (63.0%) than an increase (23.2%), and some of these changes were quite large (Table 3).

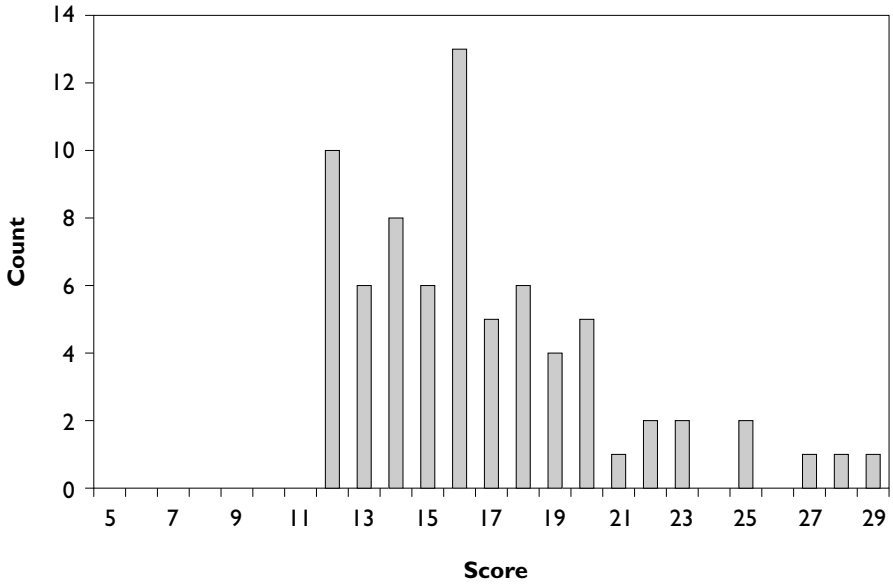

Figure 1 **Midpoint score distributions for CAST-1**

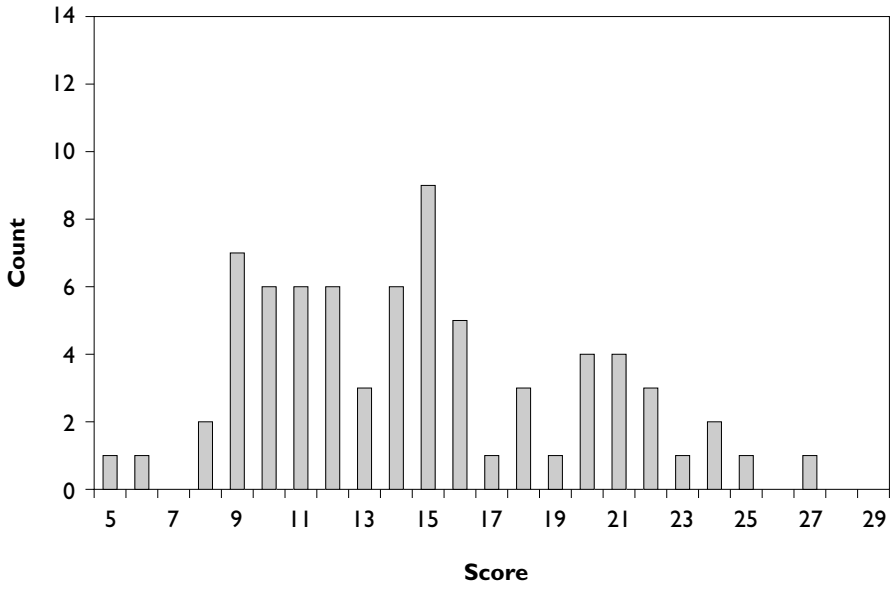

Figure 2 **Midpoint score distributions for CAST-2**

Table 3 **Score difference: CAST-2 minus CAST-1**

Score difference	*Frequency*	*%*	*Cumulative %*
–10	1	1.4	1.4
–8	2	2.7	4.1
–7	5	6.8	11.0
–6	5	6.8	17.8
–5	3	4.1	21.9
–4	8	11.0	32.9
–3	11	15.1	47.9
–2	7	9.6	57.5
–1	4	5.5	63.0
0	10	13.7	76.7
1	7	9.6	86.3
2	5	6.8	93.2
3	2	2.7	95.9
4	2	2.7	98.6
5	1	1.4	100.0
Total	73	100	100

Sensitivity analysis 1

Twenty-five (34%) of the CAST-1 questionnaires and 20 (27%) of the CAST-2 questionnaires had some missing data (Table 4). Five items from CAST-1 had between 5 and 6 percent missing data, as did two different items from CAST-2. When data were analysed treating all missing data as if the item score would have been zero (observed score), the indices of agreement did not alter. The kappa statistic for agreement across the <15 and ≥15 score groups was 0.46 ($p < 0.001$), and the overall agreement was 72.6 percent (95% CI: 61, 82). Children were still more likely to decrease in score group than to increase in score group. When scoring in three categories, the weighted kappa increased to 0.29, indicating that there was fair test–retest reliability. When using the observed score, for CAST-1 the median score was 16 (IQR 13–18, range 9–29), and for CAST-2 the median score was 14 (IQR 10–18, range 5–27). The correlation between the two test results was 0.66 (Spearman's rho), and the Wilcoxon signed-rank test still showed a significant difference between test pairs ($p = 0.001$). The median difference between scores at CAST-1 and CAST-2 was −2 (IQR −4.5–1, range −10–5).

Sensitivity analysis 2

When data were analysed treating all missing data as if the item score would have been 1 (maximum score), the indices of agreement dropped slightly. The kappa statistic for agreement across the <15 and ≥15 score groups was 0.36 ($p < 0.001$), and the overall agreement was 68 percent (95% CI: 57, 79). Children were still more likely to decrease in score group than to

Table 4 **Number of missing items on CAST-1 and CAST-2**

Number of missing items	*CAST-1*	*CAST-2*
	N (%)	*N (%)*
0	48 (65.8)	53 (72.6)
1	13 (17.8)	12 (16.4)
2	4 (5.5)	3 (4.1)
3	3 (4.1)	1 (1.4)
4	1 (1.4)	1 (1.4)
5	1 (1.4)	0
6	2 (2.7)	0
7	0	1 (1.4)
8	0	2 (2.7)
9	1 (1.4)	0
Total	73	73

increase in score group. When scoring in three categories, the weighted kappa dropped to 0.23, indicating that there was fair test–retest reliability. The median score was 16 (IQR 14–19, range 12–29), and for CAST-2 the median score was 15 (IQR 11–18.5, range 5–27). The correlation between the two test results was 0.65 (Spearman's rho), and the Wilcoxon signed-rank test still showed a significant difference between test pairs ($p = 0.001$). The median difference between scores at CAST-1 and CAST-2 was −2 (IQR −4–0, range −10–+5).

Sensitivity analysis 3

When data were analysed using the observed score at CAST-1 and the maximum score at CAST-2, the indices of agreement dropped slightly. The kappa statistic for agreement across the <15 and ≥15 score groups was 0.31 ($p < 0.007$), and the overall agreement was 66 percent (95% CI: 54, 76). Children were now no more likely to decrease in score group than to increase in score group. When scoring in three categories, the weighted kappa dropped to 0.25, indicating that there was fair test–retest reliability. The correlation between the two test results was 0.62 (Spearman's rho), and the Wilcoxon signed-rank test still showed a significant difference between test pairs ($p = 0.003$). The median difference between scores at CAST-1 and CAST-2 was −2 (IQR −4–+1.5, range −9–+5).

Discussion

Main findings

In this sample, overall there was moderate test–retest reliability between screen results using the CAST (CAST-1) and a CAST retest (CAST-2) within an average of 2 months. This was demonstrated by a kappa statistic of 0.41 (<15 versus ≥15), and the finding that 73.8 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 what level of reliability the CAST has in a high scoring sample. It was informative to discover that children were more likely to move down score groups over time.

As an extreme sample was approached in this study, it is important to consider whether the statistical phenomenon of regression to the mean could explain the significant downward movement of scores on second administration of the CAST. In this case, the percentage of regression to the mean was calculated at 33 percent,[2] indicating that regression to the mean may partly explain this downward trend. This may also be partly explained by missing data, as the effect disappears in the third sensitivity analysis.

As the 12–14 group has also been used for sampling in epidemiological research in order to assess whether any autism spectrum condition cases are found amongst those scoring below the cut-point, it was valuable to discover that the test-retest reliability across all scoring groups was fair with a weighted kappa of 0.25 across three score groups.

The score cut-points on the CAST are still provisional. It was useful to find a moderate correlation between the scores at CAST-1 and CAST-2 in the sampling group, treating the CAST as a whole scale. The Wilcoxon signed-rank test showed a possible difference between score distributions at CAST-1 and CAST-2, which may in part reflect the time lag between administrations or a developmental effect related to the acquisition of skills. Alternatively, this difference could point towards increased parental observation of their child following administration of CAST-1, and deciding that behaviours they thought their child was exhibiting were not in fact evident. When the effect of missing data was examined, the difference between score distributions was still present. It is therefore likely to be an important difference in terms of whether or not a child would be selected for further assessment in an epidemiological study.

The CAST was found to have moderate test–retest reliability, which was lower than previously published studies of autism screening tests. For example, the Autism Spectrum Disorders in Adults Screening Questionnaire (ASDASQ) (Nylander and Gillberg, 2001) reported a correlation of 0.82 (Spearman's rho), whilst on the CAST it was 0.67. However, no other published study has investigated test–retest reliability in a sample enriched by individuals with high scores, therefore making a comparison between this and other studies impossible. However, a related study examining test–retest reliability of the CAST in a population sample revealed a correlation of 0.83 (Williams et al., 2006), as well as a kappa statistic of 0.70 (<15 versus ≥ 15), indicating good test–retest reliability. It is also difficult to make further comparisons with other tests due to the different statistical measures used; however, both the Autism Spectrum Quotient (AQ) used to screen adults (Baron-Cohen et al., 2001) and the High-Functioning Autism Spectrum Screening Questionnaire (ASSQ) used to screen children and adolescents with normal intelligence or mild mental retardation (Ehlers et al., 1999) reported high correlation coefficients (0.7 and 0.96 respectively).

Strengths and weaknesses of the study and further research

A proportion of the questionnaires had missing data. The effect of missing data was examined by investigating the extremes of possible scores, the minimum and maximum that the individual could have received. The agreement and reliability indices were similar in both cases, and therefore

we can be confident that missing data did not affect the overall test–retest reliability of the CAST.

This study of test–retest reliability was carried out in a small, higher scoring sample than previously studied. It was not possible to ensure that the same parent completed both questionnaires, which may partially explain movement between scores groups as parents may disagree in their perceptions of social and communication skills in their child. Also, although the median time interval between questionnaire administrations was within 2 months, over 25 percent of CAST-2 questionnaires were completed 3 months or more after CAST-1. Furthermore, the CAST-2 was completed at the time of assessment and it was not possible to ensure that it was administered before the ADI–R. Although most participants were given the CAST-2 prior to the ADI–R, some may have been given it afterwards to complete; the lower CAST score at retest could be explained by the possibility that some parents may have benefited from taking part in the ADI–R before completing the CAST for a second time, in that after the ADI–R they had a greater understanding of the areas that the CAST questionnaire was trying to draw on.

Whilst this study points to the conclusion that the CAST near the cut-point of 15 is relatively unstable, it is useful to identify that the CAST score decreases over time. If it were the case that there was a trend towards an increased score at CAST-2, this could mean that there is a potential for an increase in the rates of false negatives, as the cut-point of 15 is used for sampling. This investigation is part of a large epidemiological study, one aim of which is to investigate the prevalence of autism spectrum conditions.

It may be useful to administer the CAST a second time for the high scoring (15+) and medium scoring (12–14) individuals prior to selecting them for assessment. This may serve to reduce the number of screen false positives and save resources. However, it is first important to establish whether children whose score at retest may have taken them below the screen cut-point later received a research diagnosis. This is currently being investigated.

Conclusion

The CAST has shown moderate test–retest reliability in a small sample of children who score over and around the cut-point of 15. These results suggest that the CAST is relatively robust for screening for autism spectrum conditions in epidemiological research. The CAST can be recommended as a screening test for autism spectrum conditions in epidemiological studies, but it is not appropriate to recommend the use of the CAST as a general population screening test in a public health or educational setting as there

is insufficient evidence regarding the effectiveness of a screening programme as a whole (National Screening Committee Child Health Subgroup, 2005). The development of this screening test, however, contributes to the body of evidence required to decide whether screening may be appropriate in the future.

Acknowledgements

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). Simon Baron-Cohen was also supported by the MRC during the period of this work. We are grateful to the schools that participated, and to the parents who gave up their time to complete the questionnaires.

Notes

1 The CAST is available in two articles (Scott et al., 2002; Williams et al., 2006) and electronically at: http://www.autismresearchcentre.com/instruments/research_instruments.asp.

2 This was estimated as $P_{rm} = 100(1 - r)$, where P_{rm} is the percentage of regression to the mean, and r is the correlation between CAST-1 and CAST-2.

References

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

AMERICAN PSYCHIATRIC ASSOCIATION (1994) *Diagnostic and Statistical Manual of Mental Disorders*, 4th edn (DSM-IV). Washington, DC: APA.

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

LORD, C., RUTTER, M. & LE COUTEUR, A. (1994) 'Autism Diagnostic Interview–Revised: A Revised Version of a Diagnostic Interview for Caregivers of Individuals with Possible Pervasive Developmental Disorders', *Journal of Autism & Developmental Disorders* 24 (5): 659–85.

LORD, C., RUTTER, M., DILAVORE, P.C. & RISI, S. (2001) *Autism Diagnostic Observation Schedule*. Los Angeles, CA: Western Psychological Services.

NATIONAL SCREENING COMMITTEE CHILD HEALTH SUBGROUP (2005) 'National Screening Committee Policy on Screening for Autism', retrieved 15 February 2006

from http://libraries.nelh.nhs.uk/screening/viewResource.asp?categoryID=1330&uri=http%3A//libraries.nelh.nhs.uk/common/resources/%3Fid%3D60310.

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) *Stat 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.

WILLIAMS, J., ALLISON, C., SCOTT, F., STOTT, C., BOLTON, P., BARON-COHEN, S. & BRAYNE, C. (2006) 'The Childhood Asperger Syndrome Test (CAST): Test–Retest Reliability', *Autism* 10 (4): 415–27.

WORLD HEALTH ORGANIZATION (1993) *The ICD-10 Classification of Mental and Behavioural Disorders: Diagnostic Criteria for Research*. Geneva: WHO.