autism © 2005
SAGE Publications
and The National
Autistic Society
Vol 9(1) 45–68; 049029
1362-3613(200502)9:1

The CAST (Childhood Asperger Syndrome Test)

Test accuracy

JO WILLIAMS *University of Cambridge, UK*

FIONA SCOTT *University of Cambridge, UK*

CAROL STOTT *University of Cambridge, UK*

CARRIE ALLISON *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 parental questionnaire to screen for autism spectrum conditions. In this validation study, the CAST was distributed to 1925 children aged 5–11 in mainstream Cambridgeshire schools. A sample of participants received a full diagnostic assessment, conducted blind to screen status. The sensitivity of the CAST, at a designated cut-point of 15, was 100 percent, the specificity was 97 percent and the positive predictive value was 50 percent, using the group's consensus diagnosis as the gold standard. The accuracy indices varied with the case definition used. The sensitivity of the accuracy statistics to case definition and to missing data was explored. The CAST is useful as a screening test for autism spectrum conditions in epidemiological research. There is not currently enough evidence to recommend the use of the CAST as a screening test within a public health screening programme in the general population.

KEYWORDS *autistic disorders; childhood developmental disorders; pervasive disorders; screening*

ADDRESS *Correspondence should be addressed to:* JO WILLIAMS, *Department of Public Health and Primary Care, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, UK. e-mail: jo.williams@phpc.cam.ac.uk*

Introduction

Screening for autism spectrum conditions may be desirable as a public health service or as a part of epidemiological research. Screening as a public health service is a means of actively identifying cases where there may or may not be a previous concern about development. It has been shown that

DOI: 10.1177/1362361305049029

the mean age of diagnosis for typical autism is 5.5 years, and as late as 11 years for Asperger syndrome, in spite of much earlier parental worries (Howlin and Moore, 1997). Screening might be able to bring the age of diagnosis earlier, and also function to reassure the worried well. Earlier diagnosis may be desirable for a number of reasons: to allow time for genetic counselling; to initiate parental support; and to allow for earlier intervention (Baird et al., 2001).

Currently there is insufficient evidence to recommend screening for autism spectrum conditions as a public health service (National Screening Committee Child Health Subgroup, 2001). One of the gaps in the evidence is the lack of a screening test that has been fully validated and shown to be effective in the general population. This article provides evidence relevant to this gap.

An effective screening test for autism spectrum conditions would also be invaluable for epidemiological research. Due to the resource implications it would not be possible to undertake a detailed assessment of all children in a large population-based study. A screening test can be used in a first phase of an epidemiological survey to sift out the children who require further detailed assessment in a second phase of the study, and hence make large studies feasible.

The focus of this study is on primary-school-age children. Potential screening tests for typical autism in preschool children have been developed (Baird et al., 2000; Robins et al., 2001). It is appropriate to develop a screening test for primary-school-age children, as many children with autism spectrum conditions are not identified prior to school entry. Coverage of preschool surveillance is incomplete, and the existence or severity of an autism spectrum condition may only become apparent in the new and demanding environment as a child enters school (Hall and Elliman, 2003).

Numerous screening tests have been written that can be used with primary-school-age children. These include: the Australian Scale for Asperger Syndrome (Atwood, 2001); the Children's Social Behaviour Questionnaire (Luteijn et al., 2000); the Pervasive Developmental Disorders Questionnaire (Baird et al., 2000); the Asperger Syndrome Screening Questionnaire (Ehlers and Gillberg, 1993; Ehlers et al., 1999); the Autism Behaviour Checklist (Krug et al., 1980); the Gilliam Autism Rating Scale (Gilliam, 1995; South et al., 2002); and the Social Communication Questionnaire (Berument et al., 1999).

There are no published validation studies available for the Australian Scale for Asperger Syndrome or the Pervasive Developmental Disorders Questionnaire. Both sensitivity and specificity estimates are not available from studies of the Children's Social Behaviour Questionnaire or the Gilliam Autism Rating Scale. The Social Communication Questionnaire has been validated in two studies (Berument et al., 1999; Bolte et al., 2000), and has demonstrated good sensitivity and specificity. However both these studies were in clinical samples, and the test needs further validation in the general population. The Asperger Syndrome Screening Questionnaire has been validated in a clinical sample (Ehlers et al., 1999) and showed good sensitivity and specificity. Whilst it has been used in the general population (Ehlers et al., 1999), data on sensitivity and specificity are not available in this context.

Many promising screening tests are being developed, but there is currently no screening test for autism spectrum conditions which has been fully validated in the general population, which has been shown to be effective, and for which information about validation is available in the public domain. The aim in further developing the Childhood Asperger Syndrome Test (CAST) was to validate a test for use in the general population rather than clinical populations, and to develop a test that is sensitive to autism spectrum conditions, including pervasive developmental disorder not other specified (PDD-NOS), not just to typical autism.

The CAST is a 37-item parental self-completion questionnaire, shown in the Appendix. There are some points to make about the name of the questionnaire. The CAST is not, strictly speaking, specific to Asperger syndrome, but it was developed to be sensitive to autism spectrum conditions in the mainstream school population, and therefore for use predominantly in children with cognitive ability within the normal range. Therefore many, though not all, of the children identified with an autism spectrum condition using the CAST will have Asperger syndrome. The name CAST is kept for the purposes of this article to maintain continuity with the test's previous publication (Scott et al., 2002a).

There is an ongoing debate over whether autism represents an extreme end of normal variation in behaviour or qualitatively different behaviours (Volkmar et al., 1997). The CAST was designed as a quantitative scale and assumes that behaviours fall on a continuous distribution, and is based on a dimensional conceptualization of autism spectrum conditions and related social and communication difficulties. It is possible, however, to impose arbitrary cut-points on the continuum to delineate categories of behaviour that are qualitatively different from normal behaviour, and the CAST is therefore compatible with a categorical conceptualization of autism.

Details of the instrument development of the CAST have been published previously (Scott et al., 2002a). Two previous pilot studies have been conducted (Scott et al., 2002a). The first pilot was in a small sample of known diagnostic status. This study demonstrated that the CAST discriminates well between children with Asperger syndrome and normally developing children. A preliminary cut-point of 15 was chosen, as all the children with a diagnosis of Asperger syndrome scored at 15 or above and none of the normally developing children scored above 15. A second pilot study was in a population-based sample of 1150 children in mainstream schools. The cut-point of 15 was used again and showed that the CAST has good specificity at this point (98 percent). The response rate in the population sample was very low (17 percent), and it was not possible to calculate the sensitivity as children with a low score on the CAST were not given a full diagnostic assessment. The aims of this article are to further validate the CAST in a larger population sample, to improve the response rate, to generate sensitivity data, and to confirm a suitable cut-point for the CAST.

Methods

School selection and response

Six schools were selected to represent different geographical areas of Cambridgeshire: two in Cambridge city, one in North Fenlands, one in East Fenlands, and two in West Fenlands. Large schools were selected for convenience. Each of the headteachers received a letter of invitation to join the study, which was followed by a meeting between each headteacher who was interested in taking part, and two members of the research team (FS, JW). The aim of this meeting was to explain further details about the study, and to provide an opportunity for the headteacher to ask questions. A training session for the staff on Asperger syndrome was offered. One of the schools took up this offer. Five of the schools agreed to take part, with one of the Cambridge city schools refusing. The percentage of children on the special needs registers of the participating schools ranged from 18 to 66 percent (mean = 34 percent, SD = 19 percent) (Ofsted, 2003).

Questionnaire distribution

Each school was asked to distribute a copy of the CAST to each child in the school who was between the ages of 5 and 11. Questionnaires were distributed to the schools on 29–31 January 2001. The schools distributed the CAST during that or the subsequent week. Each child received an envelope that contained the CAST, a covering letter, and a Freepost envelope to return the questionnaire. A total of 1925 questionnaires were distributed. A second batch of questionnaires, identical to the first, was distributed to four of the schools that agreed to take part again in order to improve the response rate. This mailing was identical to the first except for the addition of a note to ask parents not to send back the questionnaire if they had already returned the first.

Returned questionnaires were excluded if the child was not in the specified age band, if they were not at one of the schools approached, or if the questionnaire was blank or a whole page was missing. A few families returned a second questionnaire on their child following the reminder mailing, and in these cases the second questionnaire was excluded.

Data entry and cleaning for the screen

The data were entered on return of the questionnaires, keeping personal and identification data separate from the screen results. A 10 percent random sample of questionnaires was double entered to audit accuracy of the data entry. There was an agreement of 98.9 percent between the two entered sets of data, and discrepancies were checked against paper versions.

The data were cleaned, checking that each entry had a unique identifier. Single-item checks were carried out for each variable to ensure that the values entered were possible and not missing if obligatory. Within-interview checks were carried out to ensure that answers were not given randomly (e.g. all 'Yes' or alternately 'Yes' then 'No') and to check that whole pages of the questionnaire were not omitted. The data were checked in this way independently by two members of the research team (FS and JW), and a consensus decision was made over any data entry ambiguities.

Questionnaire scoring and sampling

The questionnaires were scored by unweighted addition of the endorsed scoring items. A total of between 0 and 31 could be scored. Scores were grouped into three bands: ≥ 15; 12–14; <12. A score of 15 was taken as the provisional cut-point for the screening instrument. All those scoring ≥ 15 and 12–14, and a random unstratified 5 percent sample of those scoring < 12, were invited for a detailed diagnostic assessment.

Assessments

Participants in the assessment sample were contacted by telephone to arrange the assessment. Where this was not possible, they were contacted by post. Assessments were arranged between 11 and 15 months after the screen. Due to this long time lag between screen and assessment, the screening test was administered again at the start of each assessment (CAST–R). Assessments were carried out in each participant's home.

Two instruments were used as a 'gold standard' for diagnostic assessment: the Autism Diagnostic Interview–Revised (ADI–R: Lord et al., 1994) and the Autism Diagnostic Observation Schedule–Generic (ADOS–G: Lord et al., 2000). Clinical judgement is usually considered to be the diagnostic gold standard. These instruments have the advantage over clinical diagnosis of being standardized, and their reliability and validity have been shown to be good (Lord et al., 1994; 2000). No other diagnostic tools that could have been chosen were validated with the same rigour as the ADI–R and the ADOS–G. The ADOS–G was designed to differentiate between autism, autism spectrum disorder (including PDD-NOS) and non-autism (Lord et al., 2000). The ADOS–G has also been shown to discriminate between children with pervasive developmental disorders and specific developmental disorders such as specific language impairment (Noterdaeme et al., 2002). Whilst the ADI–R and ADOS–G have often been used with strict criteria to select a conservative group of cases for genetic studies, the value of using these tools as continuous measures of the wider phenotype of autistic symptoms has been described (Lord et al., 2001).

Both the interview and observation were carried out with the interviewer blind to the CAST score. Most usually one researcher did both the interview and the observation. The order of the ADI–R and ADOS–G was not randomized due to practicalities of being able to do the interview first before the child came back from school.

Reliability of assessment

Inter-rater reliability on the ADI–R and ADOS–G was assessed. A sample of videos of interviews and observations was reviewed to come to consensus codes. The mean inter-rater reliability was calculated in two ways. First, each interviewer's code was compared with each consensus code, the mean agreement across all the codes made in each interview or observation was taken, and the mean reliability across all the assessments reviewed was calculated. For the ADI–R the inter-rater reliability across all codes was 90 percent (based on ratings on one interview), and for the ADOS–G it was 87 percent (based on ratings of eight children observed). Second, weighted kappa statistics and multi-rater kappa statistics of inter-rater reliability were calculated for the ADOS–G observations using standard linear weights (Cohen, 1968; Fleiss, 1981, pp. 225–32). A weight of 1 was used for exact agreement, 0.5 for a difference of 1 in the rating, and 0 for a difference of 2 in the rating. The mean weighted kappa for the ADOS–G ratings (based on four schedules) across all non-unique rater pairs was 0.59. The multi-rater kappa statistic was 0.54. This shows that there was moderate inter-rater reliability (Landis and Koch, 1977). Data were not available to calculate kappa statistics on the ADI–R.

Assessment outcome and case definition

A case of autism spectrum condition was defined in two ways:

1. *Assessment diagnosis*. If a child scored above the cut-point for autism or autism spectrum condition on both the ADI–R and the ADOS–G, or if they had a previous clinical diagnosis of autism, Asperger syndrome or another autism spectrum condition, they were recorded as a case of autism spectrum condition.

2. *Consensus diagnosis*. There were a number of reasons for choosing a second case definition. A case definition for wider spectrum conditions including Asperger syndrome and PDD-NOS was required, and the ADI–R only provided a cut-point for autism. Some hold the opinion that the ADI–R and ADOS–G algorithms are too stringent for inclusion of PDD-NOS. For example, one study defined the criteria for PDD-NOS as scoring above two of the three domains of the ADI–R rather than all three domains, according to the algorithm (Bishop and Norbury, 2002). Also, disagreement between the ADOS–G and the ADI–R and between these tools and previous diagnoses has been observed (Bishop and Norbury, 2002).

For these reasons some researchers have used clinical judgement, based on the results of the ADI–R and ADOS–G and using international diagnostic criteria, in order to make research diagnoses, in particular for autism spectrum conditions including PDD-NOS (e.g. Bolton et al., 1994). This approach was taken for a second case definition in this study, which was referred to as the consensus diagnosis. A child was given a consensus diagnosis if they received an assessment diagnosis or were below the cut-point (≤ 2 points) in only one of the domains covered in the algorithm on either of the instruments, and the research team agreed that they met ICD-10 research criteria (World Health Organization, 1993) for a diagnosis of atypical autism, Asperger syndrome or PDD-NOS. This judgement was made by consensus by three researchers (FS, CS, JW). In practice the subgroups of autism were not differentiated, and a research diagnosis of autism spectrum condition was given.

Referral of children to clinical services

Following the assessment, parents of children who received a research diagnosis were contacted to ask if they would like their assessment data to be passed to a clinician in the research team for possible referral into clinical services. In addition, where parents had substantial concerns about their child's development that were not related to autism spectrum conditions, they were contacted to recommend that they see their GP.

Analysis

The characteristics, as recorded in the CAST questionnaires, of responders and non-responders at the assessment stage were compared to assess whether systematic bias was introduced through non-response. In addition, those invited and not invited for assessment in the lowest score group were compared. Tests for significant differences between groups were used: Mann–Whitney test for difference between medians, unpaired t-tests for differences between means, and chi-squared tests for differences between proportions. Where numbers were small, Fisher's exact test was used. It was not possible to assess the effect of non-response to the screen on the distribution of score on the CAST as descriptors of the characteristics of non-responders were not available.

The CAST scores at the time of the screen were compared with the scores at the second administration during the assessment. If an individual moved sampling group when using their maximum score (that is, the score each individual would have if each missing item were replaced with 1) in place of their observed score, their maximum score was used. Otherwise, their observed score was used.

Indices of test accuracy (sensitivity, specificity and positive predictive value) were calculated, based on observed score on the CAST. As a two-stage sampling strategy was employed, inverse probability weighting using sampling weights were used. The weights were empirical weights defined as the inverse probability of being assessed from a particular score group, reflecting both the sampling and the response rate in each score group. Confidence intervals were calculated. Where the proportion was 100 percent, confidence intervals were calculated using the weighted count to calculate a binomial exact confidence interval. If weights had not been used, the positive and negative predictive values calculated would simply have reflected the sampling strategy that led to a proportionally higher prevalence of autism spectrum conditions in the assessment sample than would be found in the general population (Feinstein, 1977; O'Toole, 2000).

Questionnaires were not omitted from the analyses due to missing data, with the exception of the exclusion of questionnaires that were blank or had whole pages missing. Two sensitivity analyses were carried out to investigate the effect of missing responses in the CAST questionnaires. First, the analysis was rerun using the maximum score. Second, if individuals crossed over a sampling boundary (from < 12 to ≥ 12, or from < 15 to ≥ 15) when their maximum score was used rather than their observed score, the analyses were rerun excluding these people.

All analyses were carried out using STATA version 7 (StataCorp, 2001).

Results

Response rates

Response rates at each phase of the study are shown in Figure 1. Overall the response rate for the screen was 26 percent, with the response rate ranging from 20 to 33 percent across the different schools, and the standard deviation across schools was 5.4 percent. There was an inverse relationship

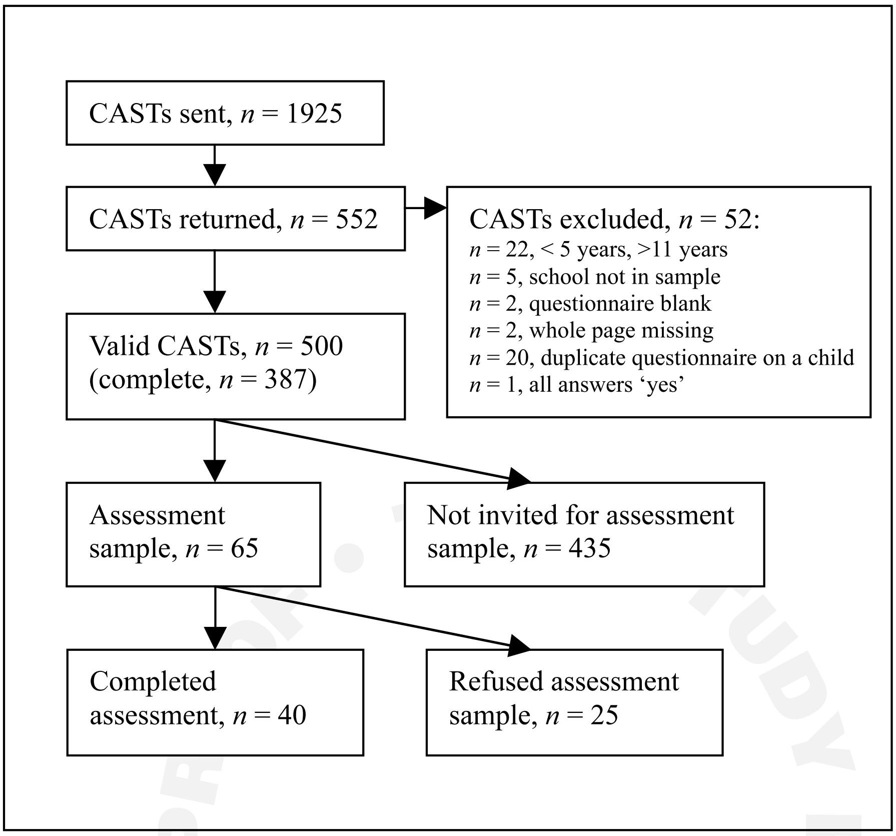
CASTs sent, n = 1925
CASTs returned, n = 552
CASTs excluded, n = 52:
n = 22, < 5 years, >11 years
n = 5, school not in sample
n = 2, questionnaire blank
n = 2, whole page missing
n = 20, duplicate questionnaire on a child
n = 1, all answers 'yes'
Valid CASTs, n = 500 (complete, n = 387)
Assessment sample, n = 65
Not invited for assessment sample, n = 435
Completed assessment, n = 40
Refused assessment sample, n = 25

Figure 1 **Response rates**

between the school response rate and the percentage of children on the special needs register, according to Ofsted reports (Ofsted, 2003). For example, the highest responding school (33 percent response) had the lowest percentage of children on the special needs register (18 percent), and the lowest responding school (20 percent) had the highest percentage of children on the special needs register (66 percent).

The response rate for the assessment was 60 percent. The characteristics of those that accepted and refused assessment are shown in Table 1. Within score groups, responders and refusers were very similar in terms of CAST score, age, gender, and parental education. Significantly more families took part where parents reported there had been concern expressed over the child's development by a teacher or a health visitor (Fisher's exact test, p = 0.017). This difference was not observed within each score group. No other differences between responders and refusers were significant.

Table 1 **A comparison of the characteristics of those who accepted and refused assessment**

Characteristic		*Group 1: < 12 on CAST*		*Group 2: 12–14 on CAST*		*Group 3: ≥ 15 on CAST*	
		Responders	*Non-responders*	*Responders*	*Non-responders*	*Responders*	*Non-responders*
Total	*N* (%)	11 (55)	9 (45)	11 (55)	9 (45)	18 (72)	7 (28)
CAST score	Median (IQR)	4 (5)	5 (4)	13 (1)	12 (1)	18 (5)	18 (4)
Age (years, decimal)	Mean (SD)	7.9 (2.0)	6.9 (1.6)	8 (2.0)	8.3 (2.1)	8.2 (1.9)	7.4 (1.8)
Gender:							
Boys	*N* (%)	4 (36)	5 (56)	5 (45)	6 (67)	15 (83)	6 (86)
Girls	*N* (%)	7 (64)	4 (44)	6 (55)	3 (33)	3 (17)	1 (14)
Age parents left education (mother, decimal years)	Mean, (SD)	17.2 (1.8)	18 (2.8)	17.2 (2.8)	16.4 (0.9)	17.8 (2.2)	17.4 (1.3)
	[missing]	[1]	[2]	[2]	[1]	[4]	[2]
Concerns expressed over child's development by teachers or health visitors[a]	Yes *N* (%)	2 (33)	0 (0)	8 (89)	4 (44)	15 (83)	7 (100)
	No *N* (%)	4 (67)	9 (100)	1 (11)	5 (56)	3 (17)	0 (0)
	[missing]	[5]	[0]	[2]	[0]	[0]	[0]
Previous diagnosis: [b]							
Language delay	Yes, no, [missing]	1, 7, [3]	0, 8, [1]	3, 5, [3]	1, 6, [2]	5, 8, [5]	2, 2, [3]
ADHD		0, 8, [3]	0, 8, [1]	1, 5, [5]	0, 6, [3]	2, 8, [8]	3, 2, [2]
Hearing/visual		1, 7, [3]	2, 6, [1]	3, 5, [3]	2, 4, [3]	4, 8, [6]	1, 3, [3]
Autism spectrum		0, 8, [3]	0, 8, [1]	0, 8, [3]	0, 6, [3]	4, 8, [6]	0, 4, [3]
Physical disability		0, 8, [3]	0, 8, [1]	0, 8, [3]	0, 6, [3]	0, 10, [8]	0, 4, [3]

[a] Significant differences between responders and non-responders (see text).
[b] Parental report on CAST at screen.

Distribution of CAST scores

The distribution of CAST scores is shown in Figure 2. Before exclusion of individuals who did not want to participate further in the research, 5.8 percent were above the cut-point of 15, and a further 4.8 percent scored from 12 to 14 on the CAST.

Thirty-nine questionnaires from the second administration of the screening test were available from the assessment sample. Only two (5 percent) had increased their CAST score so as to move up a score group. Of these, one individual moved from the lowest score group (< 12 on CAST) to the middle score group (12–14), and one individual moved from the middle score group to the highest score group (≥ 15). Twelve individuals (31 percent) moved down one sampling group, and two (5 percent) moved down two sampling groups. Twenty-three (59 percent) did not move score group.

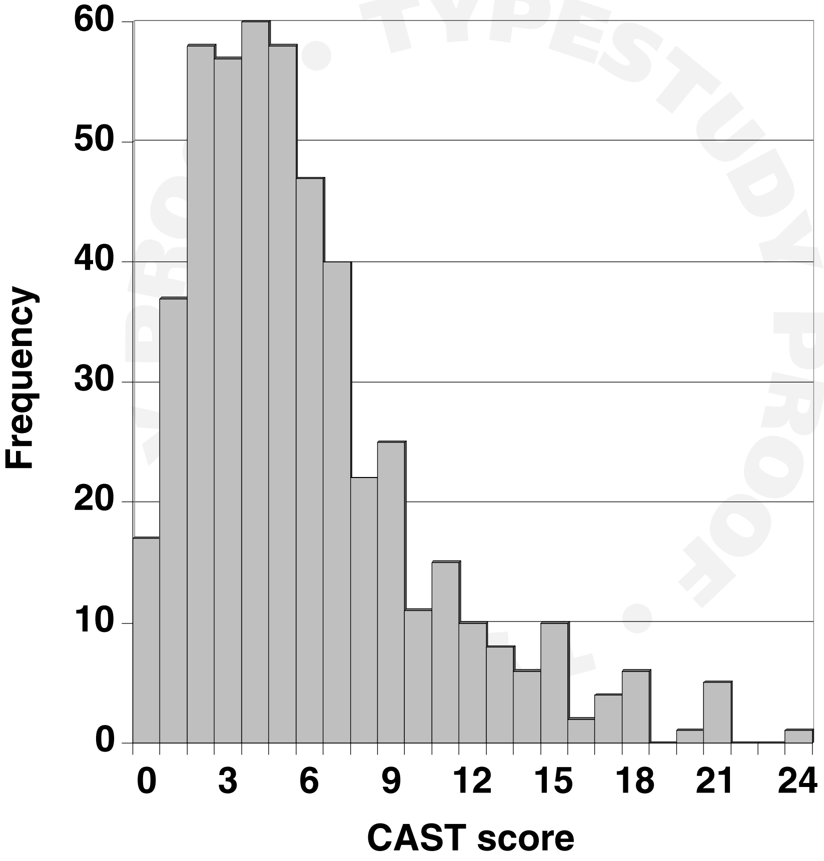

Figure 2 **Distribution of CAST scores (before exclusion of those who did not want to participate in the assessment)**

Sampling and differential verification

Those selected and those not selected for assessment were compared in order to investigate whether there may have been bias introduced by partial verification of the case status of the lowest scoring group (Table 2). There were no significant differences between those invited and those not invited for assessment.

Diagnoses at assessment

Figure 3 shows the number of cases according to the different case definitions used. Four children had previous clinical diagnoses at the time of the screen, and in addition, two children received a clinical diagnosis between the screen and the assessment. Four children were identified as

Table 2 **A comparison of the low-scoring group (< 12 on the CAST) invited for assessment against those not invited**[a]

Characteristic		*Invited*	*Not invited*
Total	*N* (%)	20	427
CAST score	Median (IQR)	4 (4)	4 (5)
Age (years, decimal)	Mean (SD)	7.5 (1.8)	7.8 (1.9)
Gender:			
Boys	*N* (%)	9 (45)	205 (48)
Girls	*N* (%)	11 (55)	222 (52)
Age parents left education	Mean, (SD)	17.5 (2.2)	17.5 (2.2)
(mother, decimal years)	[missing]	[3]	[55]
Concerns expressed over	Yes *N* (%)	2 (13)	76 (20)
child's development by	No *N* (%)	13 (87)	310 (80)
teachers or health visitors	[missing]	[5]	[41]
Previous diagnosis:			
Language delay	Yes *N* (%)	1 (6)	30 (9)
	No *N* (%)	15 (94)	312 (91)
	[missing]	[4]	[85]
ADHD	Yes *N* (%)	0 (0)	7 (2)
	No *N* (%)	16 (100)	330 (98)
	[missing]	[4]	[90]
Hearing/visual	Yes *N* (%)	3 (19)	62 (18)
	No *N* (%)	13 (81)	281 (82)
	[missing]	[4]	[84]
Autism spectrum	Yes *N* (%)	0 (0)	0 (0)
	No *N* (%)	16 (100)	334 (100)
	[missing]	[4]	[93]
Physical disability	Yes *N* (%)	0 (0)	1 (0.3)
	No *N* (%)	16 (100)	332 (99.7)
	[missing]	[4]	[94]

[a] There were no significant differences between the invited and not invited groups.

Key

—— Previous diagnosis (stated to have diagnosis at start of ADI–R)

– – Assessment diagnosis (above all cut-points on both ADOS–Gand ADI–R)

······· Consensus diagnosis

Letters refer to each case

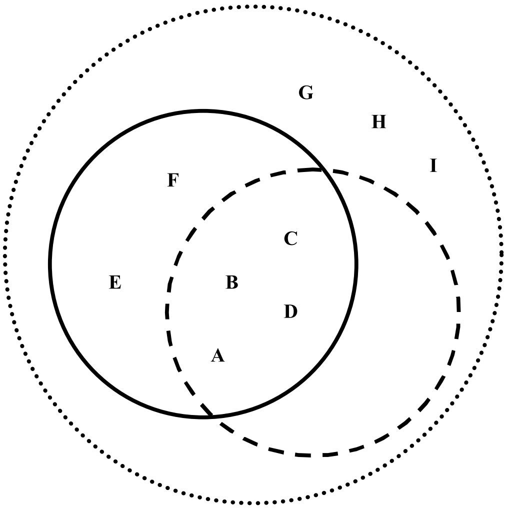

Figure 3 **Diagnoses at assessment**

cases using the assessment case definition, all of whom had a previous clinical diagnosis. However, the assessment case definition using the ADI–R and ADOS–G did not identify all the children with an existing diagnosis. A further three children were identified using the consensus case definition. The characteristics of the nine children with existing clinical diagnoses or new research diagnoses are summarized in Table 3.

Accuracy of the CAST

Figures 4a and 4b show the diagnostic accuracy of the CAST at different cut-points using the assessment diagnosis and the consensus diagnosis respectively. The consensus diagnosis captured children with wider spectrum conditions. When using the consensus diagnosis, a cut-point of 15 appeared to be appropriate where sensitivity (100 percent; 95 percent CI 74–100 percent) and specificity (97 percent; 95 percent CI 93–99 percent) were high. At higher cut-points, the sensitivity dropped. The positive predictive value was low at a cut-point of 15, at 50 percent (95 percent CI 28–72 percent). Using the assessment diagnosis, a higher

Table 3 Characteristics of participants with a previous diagnosis or a new research diagnosis

Participant	*CAST (initial screen)*		*CAST–R (retest at assessment)*		*Concerns reported in the CAST*	*Previous diagnosis*[a]	*Assessment*[b]		*Consensus*
	Score	*Max. score*	*Score*	*Max. score*			*ADOS–G*	*ADI–R*	
A	20	20	26	26	Language delay, hearing and behaviour	Y	Y	Y	Y
B	21	21	25	25	Autism spectrum, learning disabilities	Y	Y	Y	Y
C	21	21	16	16	Speech and language, autism spectrum	Y	Y	Y	Y
D	18	18	22	23	Asperger syndrome	Y	Y	Y	Y
E	18	19	25	25	Behaviour, hyperactivity	Y	N	Y	Y
F	21	21	23	23	Social interaction with peers	Y	N	N	Y
G	15	16	19	19	Social development, unusual behaviour	N	Y	N	Y
H	17	19	22	22	Hearing difficulties	N	N	Y	Y
I	17	17	14	14	Hearing difficulties, behaviour	N	Y	N	Y

[a] Previous diagnosis of autism spectrum condition at time of interview.
[b] Above cut-point on algorithm.

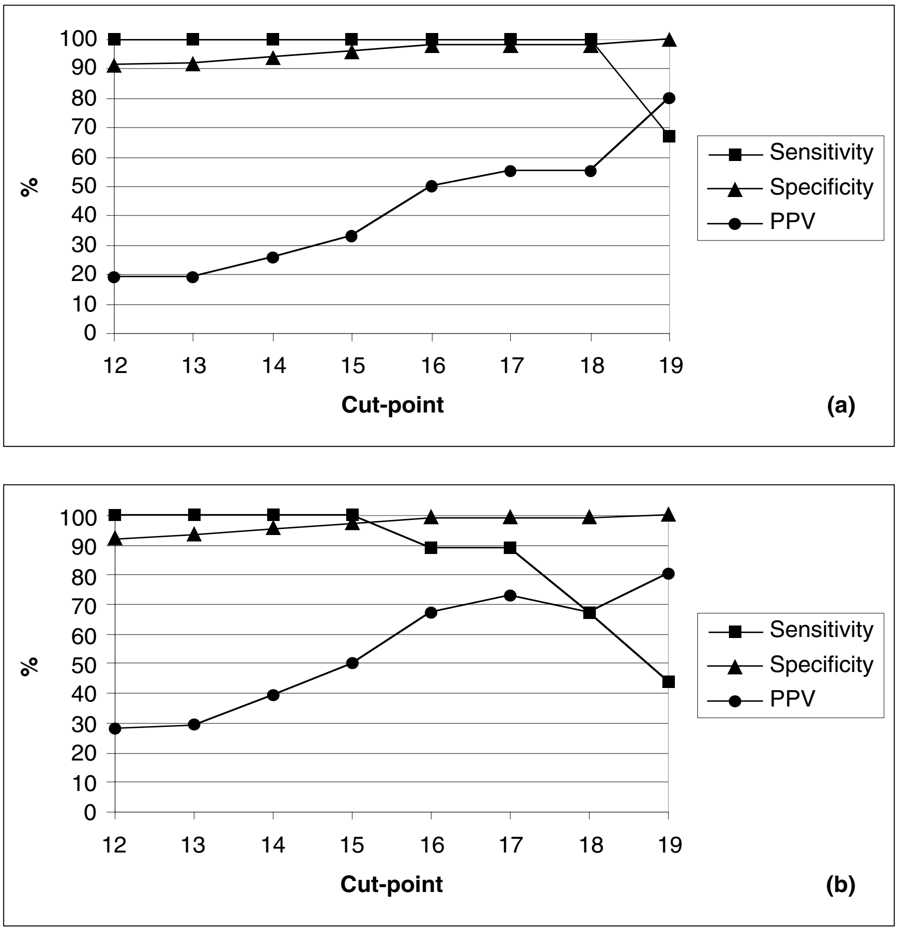

Figure 4 **Accuracy indices at each cut-point for (a) assessment diagnosis (b) consensus diagnosis**

cut-point may be more appropriate, such as 18 where sensitivity was 100 percent (95 percent CI 63–100 percent) and specificity 99 percent (95 percent CI 96–100 percent).

Missing data in the CAST: sensitivity analyses

Of the CAST questionnaires, 387 (77 percent) were complete, 85 (17 percent) had one or two missing items, 22 (4 percent) had three or four items missing and six (1 percent) had between five and nine items missing. In the first sensitivity analysis, using an individual's maximum score in place of their observed score, the sensitivity was not affected; however the specificity and positive predictive value dropped. For example, if a cut-point of 15 was applied with consensus diagnosis and an individual's maximum score was used in place of their observed score, the specificity dropped a little to 95 percent (95 percent CI 90–98 percent) and the positive predictive value dropped to 39 percent (95 percent CI 21–61 percent). Similarly, at a cut-point of 15 with assessment diagnosis the specificity dropped to 94 percent (95 percent CI 88–97 percent) and the positive predictive value to 26 percent (95 percent CI 12–48 percent) using the maximum score. The indices of accuracy were no different in the second sensitivity analyses as compared with those calculated using the observed scores.

Discussion

Implications of results: accuracy, validity and reliability of the CAST

This study demonstrates that the CAST has good accuracy for use as a screening test, with high sensitivity. The results are consistent with the pilot study, and demonstrate that the CAST has good specificity (Scott et al., 2002a). However, variation in performance depends on the type of diagnosis used as the gold standard.

The drawback of the CAST is the low positive predictive value, which is a function of low prevalence of the condition in the general population (O'Toole, 2000). There are major resource implications of assessing a large number of children who are false positives. There could be much anxiety associated with false positive screen results, as has been demonstrated with other screening tests (Marshall, 1996). A positive screen result brings uncertainty about health status (Marteau, 1994), in this case regarding the presence of a developmental disorder, until further assessment is undertaken. It should be noted, however, that a child who is a false positive for an autism spectrum condition may have another developmental problem which may be clinically important, as has been demonstrated in a study developing an autism screening test in children with developmental delay (Gray and Tonge, 2002). The characteristics of those who are false positive on the CAST merit further investigation.

This study has demonstrated the test–retest reliability of the CAST over a long time period and shown that scores rarely increase over time. However, many individuals decrease in score over time. Test–retest within a short time period will be explored further in a future study, as will internal consistency reliability. These results demonstrate that the CAST has good predictive criterion validity. Inter-parent reliability, construct and content validity will be further explored in future analyses.

Limitations on validity within the study

The sample size was small. As the condition has a low prevalence, some cases may have been missed simply by chance due to random sampling with a small sample, and may have contributed further to the low positive predictive value of the screen (O'Toole, 2000). It is important to replicate the validation of the CAST in a larger population sample.

A substantial number of questionnaires were incomplete. In future it is worth sampling using imputed maximum score rather than observed score as true score may have been underestimated.

Differential verification bias can occur if the application of the gold standard test differs according to the screen result. In this study, when the gold standard assessment was used, it was the same for children in all score groups. Whilst only a sample of low scorers was invited for assessment, the results demonstrate that the characteristics of those invited and those not invited were very similar, indicating that there was unlikely to be bias introduced by the sampling strategy. It is very unlikely that there were cases amongst the low scorers not invited for assessment as there were no cases identified in the middle or lowest scoring group in either this study or the pilot study (Scott, 2002a). A means of confirming that bias is not introduced in future studies because of not assessing all the low scorers would be to send a follow-up postal questionnaire to ask if any children have been referred to specialist services for assessment for developmental difficulties.

Refusal at the assessment phase may have introduced some bias as the proportion of parents who expressed concern over their child's development was higher in the responders than in the refusers. The effect on the indices of accuracy, however, is likely to be small as this difference was not observed when stratifying by score group. The time lag of approximately 1 year between screen and assessment might have weakened the predictive criterion validity. However, few participants increased in screen score between the screen time-point and the assessment time-point, showing that the CAST had moderately good temporal validity. It is possible that the decrease in scores in some participants over time might result in lower sensitivity indices if the scores from the second time point were available for the whole sample and were used for the analyses. The estimates of accuracy indices need to be confirmed with a shorter time lag between screen and assessment.

Whilst the assessments were conducted blind to the CAST score, due to the design of the assessment instruments it was not possible to stay completely blind to diagnostic status. The ADI–R has a question within the first 10 minutes asking for previous diagnoses and major concerns about the child's development or behaviour.

There is no absolute gold standard test for a developmental disorder, and existing standardized tests and clinical judgement through consensus meetings were chosen as the nearest approximations for a gold standard. The two types of gold standard were applied to demonstrate the sensitivity of the indices of accuracy to the case definition. It could be argued that the consensus diagnosis is a more appropriate gold standard as the ADI–R and ADOS–G algorithms have been shown to be less sensitive to subtler forms of autism spectrum conditions such as PDD-NOS (Bishop and Norbury, 2002).

It was not possible to control whether children were already receiving interventions for an autism spectrum condition at the time of the screen. Usually this would be a concern in the validation of a screening test. However, it is unlikely that such interventions would have masked the underlying difficulties in social and communication development to the extent that parents would not report them.

Generalizability of results

Spectrum bias is introduced when a screen is not validated in the same range of strengths and difficulties that it would be used to measure (Deeks, 2001). In this study the CAST was only validated in mainstream schools. It has not been validated in selective intake schools, or in special schools. The CAST has been validated in 5- to 11-year-olds, but not in younger or older children. Other than from the first pilot study, data are not yet available on how the CAST would perform in a higher-risk population, such as in a clinical setting, or in children referred to an educational psychologist.

Due to the low response rate to the screen, responders may not have been representative of the general population. For example, the response was lower from schools where there was a higher proportion of children with special needs. There is no way of adjusting for non-response in the analysis, as data on the characteristics of non-responders to the screen were not available at an individual level. However, as stated previously (Scott et al., 2002a), this may not be problematic if the CAST is used in the future as an early screen for children for whom there is existing concern, either parental, teacher or otherwise. This may be a more likely application than using the CAST in the general population, due to the fact that for reasons already highlighted, general population screening is not currently recommended.

Prevalence estimates are not presented for this population as they would be invalid due to low response at the screen phase of the study. Response bias is indicated by the fact that if prevalence estimates were generated within the respondents to the screen, the estimates would be very high and inconsistent with other population studies, indicating that the response to the screen is biased towards those with higher than average scores.

Response bias is further indicated in a comparison of the number of cases known to the schools and the number of cases identified through the questionnaire. This comparison demonstrated that our screening missed three in five known cases because many with previous diagnoses were among the non-responders. If these families had responded, the prevalence estimates would have been even higher. If there had been complete response to the screening test, proportionally more respondents with low scores would have been expected and lower prevalence estimates, comparable to those in other studies (e.g. Scott et al., 2002b), could be expected.

Recommendations for the improvement and use of the CAST

The CAST is demonstrating good sensitivity and specificity but low positive predictive value. The positive predictive value could be raised by validating the CAST in a clinical sample as the prevalence of the condition would be higher (Feinstein, 1977; O'Toole, 2000). As the aim is to develop a screen for the general population, however, a more pragmatic method of increasing the positive predictive value is preferable. It might be possible to introduce an additional phase prior to using the CAST, such as asking if the parent has concerns over the child's development. The CAST could then be used in a higher-risk population, and the positive predictive value may be considerably increased (O'Toole, 2000).

The CAST can be recommended as a screening test for autism spectrum conditions in epidemiological studies, as the low positive predictive value and subsequent false positives are unlikely to cause anxiety because a range of children from low to high scorers would be invited for further assessment. In addition, false positives may be of great interest in a research study as these children may be manifesting some symptoms also found in autism spectrum conditions. It is not appropriate, however, 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, 2001). 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

This study was funded by the Shirley Foundation and the Inge Wakehurst Memorial Trust. This work forms part of a PhD thesis entitled 'Screening for Autism Spectrum Disorders' (J.G. Williams, University of Cambridge, 2003). Jo Williams (née Johnson) held an MRC PhD studentship. Simon Baron-Cohen was funded by the MRC during the period of this work. The Autism Research Centre was established with the support of a grant from the NHS Anglia and Oxford Research and Development Programme. We thank Til Utting-Brown for doing the data entry of the CAST questionnaires, and Fiona Matthews (MRC Biostatistics Unit) for her statistical advice. We are grateful to the families and schools who took part.

Appendix 1: the CAST social and communication development questionnaire

Child's first name .... Child's surname ....

Child's date of birth __/__/____ Child's gender: male female

Child's birth order (e.g. 1st child in family) ....

Twin or single birth: ....

Parent/guardian's name: ....

Home address: ....

....

....

....

Home tel. no: .... Child's school: ....

Please read the following questions carefully, and circle the appropriate answer. All responses are confidential.

1 Does s/he join in playing games with other children easily? Yes No

2 Does s/he come up to you spontaneously for a chat? Yes No

3 Was s/he speaking by 2 years old? Yes No

4 Does s/he enjoy sports? Yes No

5 Is it important to him/her to fit in with the peer group? Yes No

6 Does s/he appear to notice unusual details that others miss? Yes No

7 Does s/he tend to take things literally? Yes No

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)? Yes No

9 Does s/he like to do things over and over again, in the same way all the time? Yes No

10 Does s/he find it easy to interact with other children? Yes No

11 Can s/he keep a two-way conversation going? Yes No

12 Can s/he read appropriately for his/her age? Yes No

13 Does s/he mostly have the same interests as his/her peers? Yes No

14 Does s/he have an interest which takes up so much time that s/he does little else? Yes No

15 Does s/he have friends, rather than just acquaintances? Yes No

16 Does s/he often bring you things s/he is interested in to show you? Yes No

17 Does s/he enjoy joking around? Yes No

18 Does s/he have difficulty understanding the rules for polite behaviour? Yes No

19 Does s/he appear to have an unusual memory for details? Yes No

20 Is his/her voice unusual (e.g. overly adult, flat, or very monotonous)? Yes No

21 Are people important to him/her? Yes No

22 Can s/he dress him/herself? Yes No

23 Is s/he good at turn-taking in conversation? Yes No

24 Does s/he play imaginatively with other children, and engage in role-play? Yes No

25 Does s/he often do or say things that are tactless or socially inappropriate? Yes No

26 Can s/he count to 50 without leaving out any numbers? Yes No

27 Does s/he make normal eye contact? Yes No

28 Does s/he have any unusual and repetitive movements? Yes No

29 Is his/her social behaviour very one-sided and always on his/her own terms? Yes No

30 Does s/he sometimes say 'you' or 's/he' when s/he means 'I'? Yes No

31 Does s/he prefer imaginative activities such as play-acting or story-telling, rather than numbers or lists of facts? Yes No

32 Does s/he sometimes lose the listener because of not explaining what s/he is talking about? Yes No

33 Can s/he ride a bicycle (even if with stabilizers)? Yes No

34 Does s/he try to impose routines on him/herself, or on others, in such a way that it causes problems? Yes No

35 Does s/he care how s/he is perceived by the rest of the group? Yes No

36 Does s/he often turn conversations to his/her favourite subject rather than following what the other person wants to talk about? Yes No

37 Does s/he have odd or unusual phrases? Yes No

Special needs section

Please complete as appropriate.

38 Have teachers/health visitors ever expressed any concerns about his/her development? Yes No

If yes, please specify: ....

....

39 Has s/he ever been diagnosed with any of the following?

Language delay Yes No

Hyperactivity/attention deficit disorder (ADHD) Yes No

Hearing or visual difficulties Yes No

Autism spectrum condition, inc. Asperger syndrome	Yes	No
A physical disability	Yes	No
Other (please specify)	Yes	No

Any other comments about your child? ....

....

References

ATWOOD, T. (2001) 'Diagnosis', in *Asperger's Syndrome*, pp. 13–27. London: Jessica Kingsley.

BAIRD, G., CHARMAN, T., BARON-COHEN, S., COX, A., SWETTENHAM, J., WHEELWRIGHT, S. & DREW, A. (2000) 'A Screening Instrument for Autism at 18 Months of Age: A 6-Year Follow-Up Study', *Journal of the American Academy of Child and Adolescent Psychiatry* 39 (6): 694–702.

BAIRD, G., CHARMAN, T., COX, A., BARON-COHEN, S., SWETTENHAM, J., WHEELWRIGHT, S. & DREW, A. (2001) 'Current Topic: Screening and Surveillance for Autism and Pervasive Developmental Disorders', *Archives of Disease in Childhood* 84 (6): 468–75.

BERUMENT, S.K., RUTTER, M., LORD, C., PICKLES, A. & BAILEY, A. (1999) 'Autism Screening Questionnaire: Diagnostic Validity', *British Journal of Psychiatry* 175: 444–51.

BISHOP, D.V. & NORBURY, C.F. (2002) 'Exploring the Borderlands of Autistic Disorder and Specific Language Impairment: A Study Using Standardized Diagnostic Instruments', *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 43 (7): 917–29.

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): 149–55.

BOLTON, P., MACDONALD, H., PICKLES, A., RIOS, P., GOODE, S., CROWSON, M., BAILEY, A. & RUTTER, M. (1994) 'A Case-Control Family History Study of Autism', *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 35 (5): 877–900.

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

DEEKS, J.J. (2001) 'Systematic Reviews of Evaluations of Diagnostic and Screening Tests,' in M. EGGER, G. DAVEY SMITH & D. ALTMAN (eds) *Systematic Reviews in Health Care: Meta-Analysis in Context*, 2nd edn, pp. 248–82. London: BMJ Publishing.

EHLERS, S. & GILLBERG, C. (1993) 'The Epidemiology of Asperger Syndrome: A Total Population Study', *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 34 (8): 1327–50.

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.

FEINSTEIN, A.R. (1977) 'On the Sensitivity, Specificity and Discrimination of Diagnostic Tests', in *Clinical Biostatistics*, pp. 214–26. St Louis, MO: Mosby.

FLEISS, J.L. (1981) *Statistical Methods for Rates and Proportions*, 2nd edn. New York: Wiley.

GILLIAM, J.E. (1995) *Gilliam Autism Rating Scale.* Austin, TX: Pro-Ed.

GRAY, K.M. & TONGE, B.J. (2002) 'Screening for Autism in Young Children with Developmental Delays', paper presented at the Inaugural World Autism Congress, Melbourne, 10–14 November.

HALL, D. & ELLIMAN, D. (2003) *Health for All Children*, 4th edn. Oxford: Oxford University Press.

HOWLIN, P. & MOORE, A. (1997) 'Diagnosis in Autism: A Survey of over 1200 Patients in the UK', *Autism* 1: 135–62.

KRUG, D.A., ARICK, J. & ALMOND, P. (1980) 'Behavior Checklist for Identifying Severely Handicapped Individuals with High Levels of Autistic Behavior', *Journal of Child Psychology and Psychiatry and Allied Disciplines* 21 (3): 221–9.

LANDIS, J.R. & KOCH, G.G. (1977) 'The Measurement of Observer Agreement for Categorical Data', *Biometrics* 33: 159–74.

LORD, C., LEVENTHAL, B.L. & COOK, E.H. JR (2001) 'Quantifying the Phenotype in Autism Spectrum Disorders', *American Journal of Medical Genetics* 105 (1): 36–8.

LORD, C., RISI, S., LAMBRECHT, L., COOK, E.H. JR, LEVENTHAL, B.L., DILAVORE, P.C., PICKLES, A. & RUTTER, M. (2000) 'The Autism Diagnostic Observation Schedule–Generic: A Standard Measure of Social and Communication Deficits Associated with the Spectrum of Autism', *Journal of Autism and Developmental Disorders* 30 (3): 205–23.

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 and Developmental Disorders* 24 (5): 659–85.

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.

MARSHALL, K.G. (1996) 'Prevention. How Much Harm? How Much Benefit? 3: Physical, Psychological and Social Harm', *Canadian Medical Association Journal* 155 (2): 169–76.

MARTEAU, T.M. (1994) 'Psychology and Screening: Narrowing the Gap between Efficacy and Effectiveness', *British Journal of Clinical Psychology* 33 (1): 1–10.

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

NOTERDAEME, M., MILDENBERGER, K., SITTER, S. & AMOROSA, H. (2002) 'Parent Information and Direct Observation in the Diagnosis of Pervasive and Specific Developmental Disorders', *Autism* 6 (2): 159–68.

OFSTED (2003) *Office for Standards in Education: Reports*. Available at http://www.ofsted.gov.uk/reports, 13 June 2003.

O'TOOLE, B.I. (2000) 'Screening for Low Prevalence Disorders', *Australian and New Zealand Journal of Psychiatry* 34 (Supplement): S39–S46.

ROBINS, D.L., FEIN, D., BARTON, M.L. & GREEN, J.A. (2001) 'The Modified Checklist for Autism in Toddlers: An Initial Study Investigating the Early Detection of Autism and Pervasive Developmental Disorders', *Journal of Autism and Developmental Disorders* 31 (2): 131–44.

SCOTT, F.J., BARON-COHEN, S., BOLTON, P. & BRAYNE, C. (2002a) 'The CAST (Childhood Asperger Syndrome Test): Preliminary Development of a UK Screen for Mainstream Primary-School-Age Children', *Autism* 6 (1): 9–31.

SCOTT, F., BARON-COHEN, S., BOLTON, P. & BRAYNE, C. (2002b) 'Brief Report: Prevalence of Autism Spectrum Conditions in Children Aged 5–11 Years in Cambridgeshire, UK', *Autism* 6 (3): 231–7.

SOUTH, M., WILLIAMS, B., MCMAHON, W., OWLEY, T., FILIPEK, P., SHERNOFF, E., CORSELLO, C., LAINHART, J., LANDA, R. & OZONOFF, S. (2002) 'Utility of the Gilliam Autism Rating Scale in Research and Clinical Populations', *Journal of Autism and Developmental Disorders* 32 (6): 593–9.

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

VOLKMAR, F., KLIN, A. & COHEN, D. (1997) 'Diagnosis and Classification of Autism and Related Conditions: Consensus and Issues', in D.J. COHEN & F.R. VOLKMAR (eds) *Handbook of Autism and Pervasive Developmental Disorders*, 2nd edn, pp. 5–40. New York: Wiley.

WORLD HEALTH ORGANIZATION (1993) *ICD-10: International Statistical Classification of Diseases and Related Health Problems*, 10th rev. Geneva: WHO.