Journal of Autism and Developmental Disorders, Vol. 29, No. 5, 1999

The Strange Stories Test: A Replication with High-Functioning Adults with Autism or Asperger Syndrome

Therese Jolliffe[1] and Simon Baron-Cohen[1]

Two groups of individuals, one with high-functioning autism and the other with Asperger syndrome were tested using Happé's Strange Stories Test of a more advanced theory of mind (Happé, 1994). This assesses the ability to interpret a nonliteral statement. Relative to normal controls who were IQ and age-matched, individuals with autism or Asperger syndrome performed less well on the task, while performing normally on a non-mentalistic control task. Individuals with autism or Asperger syndrome could provide mental state answers, but had difficulty in providing contextually appropriate mental state answers. Rather, their answers tended to concentrate on the utterance in isolation. This replicates Happé's result. Although the majority of both clinical groups provided context-inappropriate interpretations, the autism group had the greater difficulty. Results are discussed in relation to both weak central coherence and theory of mind.

KEY WORDS: Strange stories test; theory of mind; Asperger Syndrome; autism.

INTRODUCTION

Even the most able people with autism have difficulties appreciating nonliteral speech, such as indirect requests, sarcasm, jokes, and metaphorical expressions (Happé, 1993, 1994; Ozonoff & Miller, 1996; Tantam, 1992). For example, a recent study found that high-functioning individuals with autism are, paradoxically, more indirect or nonliteral in their interpretation of indirect requests than controls (Ozonoff & Miller, 1996). The authors interpreted the findings as suggesting that individuals with autism have overlearned the rule that questions beginning with "Can you . . . ?" should be interpreted in a nonliteral way. At the opposite extreme Baron-Cohen (1997) found that school age children with autism, with a mental age equivalent of 6 years, had difficulty seeing that a nonliteral reference (calling a cup "a shoe") might be a joke. In contrast, normal 3-year-old children expect a speaker's *intentions* might be to joke. Such pragmatic impairments are consistent with the well-established deficits in the development of a theory of mind in autism (Baron-Cohen, 1995; Baron-Cohen, Leslie, & Frith, 1985).

A final example of this pragmatic impairment in relation to the comprehension of nonliteral utterances was demonstrated on the Strange Stories test (Happé, 1994). Verbal adolescents and adults with autism of varying intellectual abilities were presented with a set of vignettes, or stories, about everyday situations where people say things they do not literally mean. For instance, someone in receipt of a birthday present says, "It's lovely, thank you. It's just what I wanted." This could be said either because it was lovely and they really did want it, or it could be said to spare the other person's feelings. In real life the different motivations that underlie people's utterances are distinguished by many factors. Such factors might be the preceding context, emotional expression, and the relationship between the speaker and hearer. The Strange Stories were written so that the motivation behind an utterance would generally be interpreted by normal individuals in just one way. These stories are more natural than standard theory of mind (ToM) tasks, and Happé recruited participants of different ToM abilities in order

[1] Departments of Experimental Psychology and Psychiatry, University of Cambridge, Downing Street, Cambridge, CB2 3EB, United Kingdom.

0162-3257/99/1000-0395$16.00/0

to look at their performance in relation to standard ToM tasks. She found that even the most able participants with autism tended to give context-inappropriate mental state explanations.

The study presented in this paper retests the Strange Stories, since as far as we are aware there has never been any attempt at independent replication. These stories were reduced in number and one was modified (with permission from Happé) in an attempt to make them even more likely to be interpreted in one way rather than another. Furthermore, a control task was presented so as to test the understanding of physical events and check the generality of any comprehension deficit that might emerge irrespective of story content. A few of these were also modified (again with permission from the author) in an attempt to make them more likely to be interpreted in one way rather than another. These control stories were more demanding than the control stories employed in the original version of the Strange Stories test, since the original version resulted in ceiling effects.

The study thus aimed to test if Happé's finding of a tendency to give context-inappropriate mental state explanations would replicate. It also sought to extend her 1994 study. Thus all participants in our study were selected for passing standard second-order ToM tasks, and were separated according to whether they had received a diagnosis of autism or Asperger syndrome. This was to test whether any weakness in processing such stories was a function of early language development, since the key difference between autism and Asperger syndrome is that in the former there is often a history of language delay whereas in the latter there is no clinically significant language delay. Second, both clinical groups were very high-functioning, having an IQ (Full-scale, Verbal, and Performance) which was at least average. Third, a normal control group matched on age, sex, handedness, and IQ (Full-scale, Verbal, and Performance) was included, since the original study did not match participants in these respects.

If the clinical groups have mentalizing difficulties (Baron-Cohen *et al.*, 1985) one would expect them not only to have difficulty providing mental state answers but to produce less of these. On the other hand if individuals with an autism spectrum disorder have weak "central coherence" (Frith, 1989) they should show evidence of this in their failure to use the context to provide the context-appropriate interpretation of an utterance. It was predicted that the individuals with autism and Asperger syndrome would show a failure to give the context-appropriate interpretation of an utterance.

Participants

The last two decades has seen a great deal of dispute about whether or not there is a single condition of autism which varies in severity, or whether there are different types which lie along a continuum. Moreover, the last decade has seen an increase in the use of the label Asperger syndrome, which diagnostically requires (among other requirements) no clinically significant delay in early language development (ICD-10; World Health Organization, 1992). Thus single words have to be used by 2 years and communicative phrases by 3 years. Pragmatics is not included. In this respect this type of autism can be distinguished from Kanner's (1943) classical cases, all of whom were clinically delayed in early language development.

Recently there has been increasing interest in how one defines autism, and how diagnostic systems should be interpreted. The DSM-IV (American Psychiatric Association [APA], 1994) suggests that one can have autism irrespective of whether or not one is delayed in early language development. There is even a question over whether Asperger's cases had Asperger syndrome (Miller & Ozonoff, 1996). The present study does not aim to contribute to this debate. Rather, we chose to distinguish individuals with a history of autism on the basis of their early language development. This gave rise to two groups one of which clinicians regarded as meeting the criteria for Asperger syndrome (ICD-10) and the other which clinicans regarded as meeting the criteria for autism (DSM-IV). We did not attempt to distinguish individuals on any other factor (e.g., motor clumsiness). Whereas our cases of autism resembled Kanner's classical cases, their symptoms had lessened which made them appear to function similarly to those with Asperger syndrome and which would have put them in the residual category according to the DSM-III-R (APA, 1987).

Our clinical participants were recruited nationally. Some were located via support groups or via letters sent out by the National Autistic Society. The majority were located through clinicians. The majority of clinical participants were or had been patients at the Maudsley Hospital (London) where experienced clinicians had determined the individuals' diagnostic status. Some were or had been patients at Charing Cross Hospital (London) where again an experienced clinician had made the diagnosis. The remainder of participants came from Elliot House, the National Autistic Society's UK center for the diagnosis of social and communication disorders.

Irrespective of whether or not an individual was attending a hospital, clinicians were contacted (with the

patients' and parents' consent) and medical and psychological or psychiatric reports were inspected. Individuals were excluded in cases where either the clinican or the parents could not be sure about early language development. As part of the checking process parents were given the revised Howlin (1995) screening questionnaire to complete. (This, devised at the Maudsley Hospital, makes use of DSM-IV criteria and seeks to identify the presence of autistic symptomatology and whether there was a clinically significant delay in early language development.) In all but one case clinicians and parents agreed on early language development. The one case where there was disagreement was excluded. Individuals whose early language development was thought to be on the borderline of what is clinically normal and abnormal were also excluded. In our attempt to be rigorous about early language development we rejected a dozen participants. There was thus no doubt about the early language development of any of our participants.

The Asperger group (i.e., the group with single words by 2 years and phrase speech by 3) contained many individuals who had not received a diagnosis until adulthood. The remainder were diagnosed in childhood or adolescence. For the former group clinicians made their diagnosis retrospectively. For the latter group, in the case of a couple of individuals, clinicians revised their diagnosis from autism to that of Asperger syndrome on the basis that these individuals did not have a clinically significant delay in early language development. According to the DSM-IV, however, some of the Asperger participants might be regarded as having autism without language delay. We have not, however, questioned the clinicians' diagnosis of Asperger syndrome, since from our point of view this leaves unaffected the key differentiator between the two groups; that of early language development. Thus in the autism group all were clinically delayed in language development and would therefore have been considered to have had a history of classical autism. In the Asperger group, none were clinically delayed in their language development.

Fifty-one adults participated in the experiment. These comprised 17 with autism, 17 with Asperger syndrome, and 17 normal adult control participants. The normal adults acted as a comparison group for the two clinical groups. The majority of clinical participants were tested in their place of residence, except where some preferred to be tested at the university. All control participants were tested in a quiet room at the university.

The 17 control participants were taken from the general population of Cambridge. These control participants were chosen to match the clinical groups as closely as possible with respect to the characteristics of age, IQ, sex, and handedness. Table I gives the participant details of chronological age (CA) verbal IQ (VIQ), performance IQ (PIQ), and full-scale IQ (FSIQ). Four one-way ANOVAs revealed no significant-differences between groups on any of these variables: CA, $F(2, 48) = 0.59$, $p = .56$; VIQ, $F(2, 48) = 0.51$, $p = .60$; PIQ, $F(2, 48) = 0.58$, $p = .57$, and FSIQ, $F(2, 48) = 0.10$, $p = .91$. The sex ratio in all three groups was 15:2 (m:f), reflecting the sex ratio found in these clinical groups in other studies (Klin, Volkmar, Sparrow, Cicchetti, & Rourke, 1995; Wing, 1981). The groups were closely matched on handedness: 15 right-handed and 2 left-handed individuals in the normal and high-functioning autism group, and 14 right-handed and 3 left-handed in the Asperger group. All participants were born in England and English was their first language. All three groups contained participants from various socioeconomic backgrounds and the three groups were broadly equivalent in terms of educational attainment. Several individuals within each group were either studying for or holding formal qualifications such as a university degree or diploma.

All participants were required to be of at least normal intelligence (i.e., scoring ≥ 85) on the WAIS-R (Wechsler, 1981, Full-scale, Performance, and Verbal IQ). All three groups prior to their recruitment were screened to check whether they had any history of psychiatric disorder, neurological disorder, or a head injury. Individuals were excluded if they reported any of these factors, and for the clinical groups parents and professionals were consulted. All participants were also required to be medication-free at the time of testing. There were also screening criteria specific to to the clin-

Table I. Participant Characteristics[a]

Participant group[a]	CA	VIQ	PIQ	FSIQ
Normal				
M	30.00	106.47	105.24	106.35
SD	9.12	10.94	14.00	12.72
Range	18–49	87–127	85–134	88–133
Autism				
M	30.71	107.59	101.77	105.12
SD	7.84	14.37	13.06	13.47
Range	19–46	88–135	85–132	90–133
Asperger				
M	27.77	110.82	100.29	107.12
SD	7.81	13.51	14.23	14.34
Range	18–49	89–130	85–133	86–132

[a] $n = 17$ in each group.

ical and control groups. The control group had to be free of any family history of autism or Asperger syndrome. The clinical groups were selected on the basis of their ability to pass both first- and second-order belief tasks[2] and they were screened to exclude those who might be depressed, since depression is much more common in autism and Asperger syndrome and can also affect social functioning and judgment.

Materials

The stimuli presented consisted of short stories. There were 18 mentalistic stories and 6 physical control stories. There were 9 types of mentalistic story, the test containing two examples of each. The 9 story-types comprised Double Bluff, Figure of Speech, Joke, Lie, Misunderstanding, Persuade, Pretend, Sarcasm, and White Lie. The 6 physical control stories did not involve mental states, nor were they social in nature. However, they did require participants to make global inferences that went beyond what was explicitly mentioned in the text.

The mentalistic stories contained two and, for one story type, three test questions; the comprehension question, which usually took the form, "Was it true what X said?" and the justification question(s) which usually took the form, "Why did X say that?" The physical control stories had just one question, asking why a particular action had been carried out.

Each of the stories with their questions appeared on white A5 sheets. The mentalistic test stories also contained a small black and white line drawing of the significant characters mentioned in the story. These line drawings were simple, but showed emotional expressions and illustrated the contextual setting. All the stories were laminated. All stories had an identifying word label that appeared in the top right-hand corner, the purpose of which was to facilitate scoring.

The stories that were used in this experiment were taken from Happé (1994) and a few were excluded and amended with permission from the author. Examples of the mentalistic and physical stories can be found in the Appendix.

[2] Participants were given first- and second-order theory of mind tests. The first-order task was a version of Perner, Frith, Leslie, and Leekam's (1989) Smarties task. The second-order task was Baron-Cohen's (1989) ice cream van test. Whereas all participants passed the first-order task, 5 out of 51 participants failed the second-order task. These included 1 participant with Asperger syndrome, 2 with high-functioning autism, and 2 normal control participants. These participants were retested on a new variation of the second-order belief task and all were found to pass.

Procedure

Each participant was tested by themselves in a room which was free from distractions. The experimenter sat next to the individual so that she could read them the stories.

The Test Items

The mentalistic task was always presented first, followed by the physical control task. Since the control stories always came after the mentalistic stories, the latter acted as a control for fatigue and motivational deficits, as well as being a test of nonsocial global inferences.

The experimenter shuffled the mentalistic stories thoroughly before they were given to each participant. These stories were therefore presented in a different random order to each individual. This was done to ensure that any significant effects were not due to story order and, as a further safeguard, the experimenter made sure that the two examples of each type of story never appeared together.

The stories being presented were placed face up on the table directly in front of the participant. The participant was told that he/she were going to be read some stories. They were instructed to listen carefully as they would be required to answer questions about the stories.

After each of the mentalistic stories had been read, the experimenter asked the comprehension and mental state question(s). The story remained in front of the participant not only throughout the reading but also throughout questioning. This was done in order to minimize memory requirements. The comprehension question usually took the form, "Was it true what X said?" On the very rare occasions when a participant made an error, the story was read out again, until the participant answered correctly or justified their answer and appeared to understand (e.g., "well no, it's not literally true, but it is the correct expression to use"). The mental state question(s), usually took the form, "Why did X say that?" The participants were given as long as they felt necessary and were encouraged to read the story for themselves a second and even a third time in order to facilitate their explanation for the story character's utterance.

After the mentalistic stories had been completed, participants were given the physical control stories. These stories were read out loud to each participant, and after each one had been read, its inference question was asked, asking why something happened or why a particular action had been carried out. As with the mentalistic stories, the stories remained in view while the inference question was being asked. This was again done to minimize memory requirements. Furthermore,

to be consistent with the mentalistic stories, participants were given as long as they felt necessary to come up with the answer, and they were encouraged to read the story through themselves in order to facilitate their explanation of why something had happened or why a particular action was carried out.

Scoring

For the mentalistic stories participants were scored on their answers to both the comprehension and "Why" questions. For the comprehension question, participants' original responses were noted, along with any amendments. For the "Why" question, scoring was a little more complicated. Justifications could be correct or incorrect. But a correct justification could be correct in two ways: it could be a correct physical explanation or a correct mental explanation. Similarly, an incorrect justification could be incorrect in two ways, it could be an incorrect physical explanation or an incorrect mental explanation. Thus in the story where Emma is playing and pretends that a banana is a telephone, Emma's statement that the banana is a telephone can be correctly justified in two ways: in the physical sense, "because the banana is shaped like a telephone," and in the mental sense, "because Emma is pretending that the banana is a telephone." Similarly, an incorrect justification could be incorrect in two ways: in the physical sense, an answer that is factually incorrect, such as "because she is about to eat the banana" and in the mental sense, "because Emma is joking."

Where both a correct and incorrect justification were given, participants were scored on their correct justification. Thus participants were given credit for their best answer. Similarly, if a participant's answer appealed to both physical and mental states, the justification was scored as a mental state.

Physical state answers included terms such as big, looks like, is shaped like, to sell them, to get rid of them, to not get X (physical outcome, e.g., a filling, told off, mugged). Mental state answers included all those that referred to thoughts, feelings, desires, traits, and dispositions. Mental state justifications included terms such as like, want, think, know, happy, cross, joke, pretend, lie, afraid, please, hurt, expect, and to fool.

For the physical control stories, participants were scored on their answers to the global inference question, which asked why something happened or why a particular action had been carried out. On the rare occasions where an individual gave more than one explanation for why a particular action was carried out, or amended their answer, it was always their best answer that was accepted. Accenting amendments to answers, or giving credit for the participant's best answer, was to parallel and be consistent with the scoring of the mentalistic stories.

The participant's responses were noted, including any amendments, next to the story's identification label which appeared on each score sheet. The participant's answers were recorded in full on the score sheet so that they could be analyzed at a later date. There were just a few instances where participants were unable to give a response to the "Why" question of the mentalistic stories, and to the action question of the physical control stories. These omissions are examined in the Results section.

Judging whether an explanation is appropriate or not is clearly subjective. However, the stories were selected and amended to ensure that they were unambiguous and therefore only one explanation was reasonably appropriate for both the mentalistic stories and for the justification of why a particular action occurred. Although the stimuli were relatively unambiguous, there is still a degree of subjectivity surrounding participants' statements, therefore validation of the scoring was undertaken in order to establish its validity. All of the statements made by 9 of the participants in each group were given to a second rater, who was blind to the identity and diagnostic status of the participants and naive to the hypothesis being tested. The degree of concordance was 98% for each of the two conditions.

RESULTS

The first performance measure was aimed at assessing whether the groups differed in their ability to provide context-appropriate explanations (justifications) for a story character's nonliteral utterance in the mentalistic condition and whether they could make an inference or give a reasoned explanation as to why a particular action had occurred in the physical condition. The performance scores on these two conditions were measured in percentages (so as to equate the mentalistic and physical tasks which had 18 stimuli in the former and 6 in the latter). Therefore individuals could obtain scores ranging from 0% to 100% for each of the conditions. The three groups mean percentage correct for their explanations on the mentalistic and physical stories are illustrated in Table II.

Accuracy on the Mentalistic and Physical Stories

The accuracy scores were approximately normally distributed and the variances were approximately equal,

Table II. Percentage of Justifications that were Correct on each Condition

Participant group[a]	Mental	Physical
Normal		
M	99.67	88.18
SD	1.35	12.81
Range	94–100	67–100
Autism		
M	84.31	84.31
SD	10.80	10.80
Range	56–94	56–94
Asperger		
M	89.22	87.18
SD	9.31	12.49
Range	72–100	67–100

[a] $n = 17$ for each group.

according to Cochrans C test, so a two-factor repeated-measures ANOVA was performed on the mean percentage scores for each of the three groups. This ANOVA had a between-participant variable of Group, and a within-participant variable of Condition (Mentalistic and Physical). The ANOVA revealed a significant main effect of Group, $F(2, 48) = 4.66$, $p < .05$; and Condition, $F(1, 48) = 12.02$, $p < .001$, which suggests that the groups differed in the overall correctness of their statements and that the type of condition (whether mentalistic or physical), had some effect on performance. The higher order interaction of Group × Condition, was also, as predicted, significant, $F(2, 48) = 7.41$, $p < .01$.

Given that the Group effect was significant, the Group × Condition interaction was investigated further to see if there were different Group effects for the two conditions. To examine whether this group effect applied to one or both of the conditions, simple effects were examined which compared the different Groups on each Condition. As predicted analysis of simple effects showed the effect of Group to be significant only for the Mentalistic condition, $F_{\text{Mentalistic}}(2, 48) = 15.30$, $p < .001$; $F_{\text{Physical}}(2, 48) = 047$, $p = .63$.

The source of the Group × Condition interaction was investigated further using *t* tests. To reduce the familywise error rate only preplanned comparisons were explored. Planned contrasts of the cell means indicated that the mean percentage for the autism and Asperger groups' Mentalistic condition, were as predicted, significantly different from that of the normal control group, $t_{\text{aut.}}(48) = 5.42$, $p < .001$; $t_{\text{Asp.}}(48) = 3.69$, $p = .001$, and that the mean percentage for the autism group's Mentalistic condition had a nonsignificant trend to be lower than that of the Asperger group, $t(48) = 1.73$, $p = .09$. Observation of the mean percentages (see Table II and Figure 1) show the clinical groups to be significantly *worse* in giving context-appropriate explanations for a speaker's utterance.

Given that the Condition effect was significant it was useful to see whether there were different Condition effects for the three Groups. Simple effects were examined which compared the two conditions for each group. Analysis of simple effects showed the effect of Condition to be significant for the normal control group, $F(1, 48) = 26.01$, $p < .001$, but not the two clinical

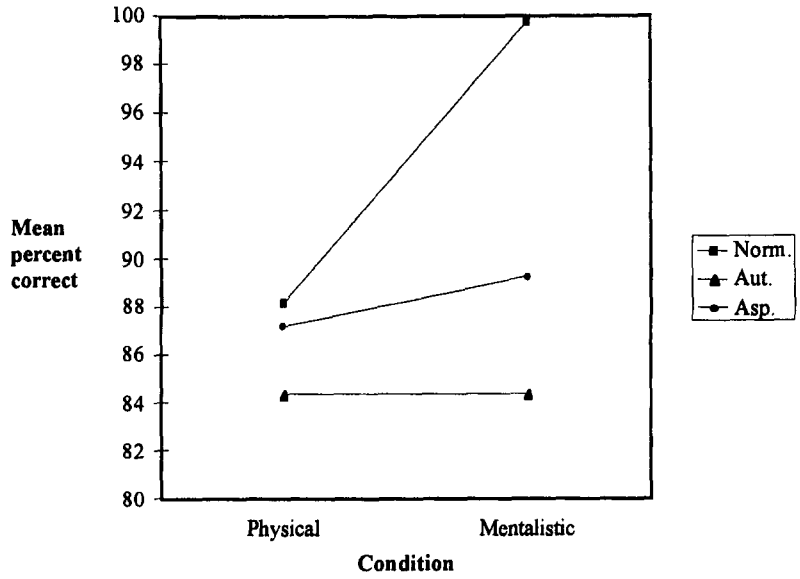


Fig. 1. Effect of Condition on justification accuracy.

groups [$F_{aut.}(1, 48) = 0.00$, $p = 1.00$; $F_{Asp.}(1, 48) = 0.82$, $p = .37$]. Observation of the means (see Table II and Figure 1) show the normal control group to be significantly *better* on the Mentalistic condition relative to their own performance on the Physical condition, whereas the autism and Asperger groups showed no differential pattern of performance on the two conditions.

Nature of Performance on the Mentalistic Stories

The next performance measure assessed the nature of participants' performance on the Mentalistic condition. The justifications on the mentalistic stories were not just correct or incorrect as was previously examined, they could in fact correspond to several types of answer. The justifications could be mentalistic or physical; appealing to mental or physical states. They could be either a correct or incorrect mental answer, or a correct or incorrect physical answer. The scores for the three groups were examined to see whether there were any differences in the types of justification provided. The unit of measurement for the type of statement was the number of answers which corresponded to the type of statement. Therefore individuals could in theory obtain scores which ranged from 0 to 18 for each of the justification types. The mean results for the three groups can be seen in Table III. The number of participants giving at least one incorrect mental state justification per story type can be seen in Table IV.

To examine whether there were any group differences for each of the justification types, a series of one-way ANOVAs were performed on the following justification types: Mentalistic Statements, Physical Statements, Correct Physical Statements, Incorrect Physical Statements, and Correct Mental Statements. Due to unequal variances, Incorrect Mental Statements had to be investigated with t tests using a separate variance estimate rather than a pooled variance estimate. The analysis revealed no difference between groups on the number of mental state justifications $F(2, 48) = 0.37$, $p = .69$; the number of physical state justifications, $F(2, 48) = 0.18$, $p = .84$; the number of correct physical state justifications, $F(2, 48) = 0.06$, $p = .94$; and the number of incorrect physical state justifications, $F(2, 48) = 1.56$, $p = .22$. However, the analysis did reveal a difference between groups on the number of correct mental state justifications, $F(2, 48) = 5.31$, $p < .01$; the source of the effect was investigated further using t tests. The analysis revealed that the clinical groups made significantly fewer correct mental state justifications than the normal control group, $t_{aut.}(32) = -3.65$, $p = .001$; $t_{Asp.}(32) = -2.18$, $p < .05$; although the clinical groups themselves did not differ, $t(32) = 0.86$, $p = 0.40$. T tests also revealed a significant difference between groups on the number of incorrect mental state justifications. The clinical groups made significantly more incorrect mental state justifications than the normal control group, $t_{aut.}(147) = -4.70$, $p < .001$; $t_{Asp.}(17.25) = -4.46$, $p < .001$, although again the clinical groups themselves did not differ significantly, $t(257) = 1.65$, $p = .11$.

To determine whether the clinical groups' tendency to provide context-inappropriate mental state answers is determined by just a few individuals in each group or was more widespread among the participants, the number of participants in each group scoring above (and below) the control group mean was calculated. This was compared to the numbers of participants in the normal group scoring above (and below) their mean.

Table III. Number of each type of Justification on the Mentalistic Condition

Participant group[a]			Correct justifications		Incorrect justifications	
	Mental	Physical	Mental	Physical	Mental	Physical
Normal						
M	14.35	3.65	14.29	3.65	0.06	0.00
SD	1.97	1.97	1.96	1.97	0.24	0.00
Range	10–18	0–8	10–18	0–8	0–1	0–0
Autism						
M	13.88	3.88	11.53	3.77	2.35	0.12
SD	1.97	1.76	2.43	1.72	2.00	0.33
Range	10–17	1–8	6–15	1–8	1–8	0–1
Asperger						
M	13.77	4.06	12.35	3.88	1.41	0.18
SD	2.39	2.28	3.10	2.15	1.23	0.39
Range	10–18	0–7	8–18	0–7	0–4	0–1

[a] $n = 17$ in each group.

Table IV. Participants giving at least one incorrect Mental State Justification

	Participant group[a]		
	Normal	Autism	Asperger
Double bluff	0	8	6
Figure of speech	0	2	1
Joke	0	4	0
Lie	0	3	2
Misunderstanding	1	5	3
Persuasion	0	6	3
Pretend	0	2	2
Sarcasm	0	6	4
White lie	0	2	1

[a] $n = 17$ in each group.

The analysis revealed that the clinical groups differed significantly from the normal control group, $\chi^2_{aut.}(1) = 30.22$, $p < .001$; $\chi^2_{Asp.}(1) = 15.07$, $p < .001$.

Comprehension and Omissions on the Mentalistic Stories

Performance on the comprehension question was examined. The number of comprehension errors were calculated for each group. Three 2×2 chi-squared tests were conducted to see whether the groups differed in their number of comprehension errors. Chi-square analysis revealed no significant difference between the three participant groups (Fisher's correction for expected frequencies < 5, $p > .05$ for all three groups).

Since only one participant in each of the clinical groups made an omission, it was thought unnecessary, given the large number of stimuli employed, to test for group differences, as omissions could not be considered a factor underlying the relatively poor performance of the clinical groups.

DISCUSSION

This study aimed to replicate Happé's (1994) main finding on the Strange Stories. Our results replicate her main findings. Thus, on the Mentalistic condition, the clinical groups were equivalent to the control group in the number of mental state justifications used. What distinguished the clinical participants on this condition was not a failure to use mental state terms (including emotion terms like "hurt") but a failure to use the appropriate mental state term for the story's context. In comparison to the normal control group the clinical groups made significantly more context-inappropriate mental state justifications, and significantly fewer context-appropriate mental state justifications. Although the clinical groups themselves did not differ in their correctness or otherwise of their mental state answers, the autism group always performed at a level below that of the Asperger group (see Tables II, III, and IV). Although every one of the participants with autism gave at least one (and on average two to three) context-inappropriate mental state answers, 12 out of 17 of the participants with Asperger syndrome gave at least one (and on average one to two) context-inappropriate mental state answers, whereas only one of the normal adults made such a mistake. The errors made by clinical participants were striking: in the sarcasm story one participant said that the lady's statement that it was a lovely day for a picnic was her "pretending that everything was OK in order to make Tom feel happier," and another participant explained the utterance in the pretend story as "a joke." As predicted, both groups of clinical participants performed normally on the Physical condition, where they were required to justify why a particular action had occurred.

Analysis of the type of justifications made on the Mentalistic condition revealed that the groups did not differ in the number of physical state justifications and the correctness or incorrectness of these. The finding that the clinical groups did not differ from the normal control group on the number of *correct* physical state justifications is at odds with Happé's finding, where her second-order (ToM) autism group made more correct physical state justifications than her normal adult control group. However, there are a few factors that might explain this different result: (a) the clinical groups recruited to take part in the experiment reported here tended to be older and intellectually more able than Happé's second-order ToM autism group. (b) Our normal control group were arguably less intelligent than those in Happé's study because there was not such a high proportion of students. (c) Some of the original mentalistic stories were excluded and one was amended.

The clinical groups' difficulties were not likely to be because they could not comprehend the stories, since they did not differ in the number of errors made on the comprehension question. Neither did the clinical groups make more incorrect physical state justifications, which is what one would have expected to see had there been general comprehension problems. Furthermore, they performed normally on the Physical condition or control task which suggests they can comprehend stories.

It is pertinent to explore why the clinical groups performed normally on the mentalistic comprehension

questions, since these can be viewed as a test of appreciating nonliteral utterances, and it is well known in the pragmatics literature that even high-functioning individuals with autism have problems appreciating such utterances (Happé, 1991, 1993, 1995; Ozonoff & Miller, 1996; Rumsey & Hanahan, 1990; Tantam, 1991). The explanation for this discrepancy seems to be due to the type of approach to or type of question asked about such statements. In the Strange Stories test there are two questions about each of the nonliteral statements. The first is the Comprehension question, which asks whether the character's statement is true, and the second is the Justification question, which asks why the character made the statement. It seems that these two questions require two levels of interpretation; the first requires a lower level of interpretation and the second requires a higher level of interpretation. Thus to answer the first question participants simply have to detect that the character's statement is at odds with the content of the story. Whereas with the second question, participants have to integrate the character's statement with the story context and therefore provide a contextually appropriate explanation. It seems that the clinical participants had no difficulty in detecting that the statement was at odds with the situation, but did have difficulty in giving a contextually appropriate explantion for why the character said what they did.

The difference between comprehension, and the integration of information for higher-level meaning, has been noted elsewhere in the autism literature. Thus Rumsey and Hamburger (1988) gave high-functioning autistic adults the Verbal Absurdity and Problem Situations of the Stanford-Binet Intelligence Test (Terman & Merrill, 1973). These authors stated that the responses to the Binet Verbal Absurdity and Problem Situation items reflected comprehension of the linguistic aspects of these problems but failures to integrate the information. Their participants, while comprehending the information, tended to provide incorrect inferences. An example is the following Problem Situation: "Helen heard a big 'Bang' and came running outdoors. There were nails all over the road, and an automobile had just stopped beside the road. What was the bang?" Whereas the correct answer was that a tire had blown, wrong answers included: an explosion occurred; it was a bomb or sticks of dynamite; a truck with nails had a crash. Rumsey and Hamburger's (1988) evidence of intact comprehension in the face of a failure to integrate information to make an appropriate inference is consistent not only with the evidence from the Strange Stories test but also with the findings from other pragmatic measures such as the finding of an impaired ability to select contextually appropriate inferences on the Test of Language Competence (Minshew, Goldstein, Muenz, & Payton, 1992).

In the Strange Stories test the clinical groups' difficulties do not seem to be due to any tendency to be less willing to make a response in comparison to their normal control group (i.e., a tendency to make more omissions) since only one participant in each of the clinical groups made an omission. Furthermore, whereas omissions could give rise to fewer *context-appropriate* answers, they could not give rise to more *context-inappropriate* answers.

The fact that the clinical groups did not fail to provide answers suggests that they did not have a problem in generating responses. Similarly, the fact that the clinical groups did not differ on the number of mental state answers provided suggests that they did not have a problem in providing mentalistic answers. Instead the problem for the clinical groups seemed to be one of providing context-appropriate mentalistic answers. This suggests a weakness in processing mental state information in context. Context-inappropriate mental state answers suggest that weak central coherence might explain the specific pattern seen on this test.

Where the clinical groups failed to use or extract meaning from the story context they tended to focus on the utterance in isolation. This resulted in them tending to generate a locally coherent rather than globally coherent answer. So, for example, a participant with autism who explains the white lie as a joke may be failing to use the story context to inform his answer. Thus, this individual would be making an inference about how a character could have felt but not how he/she actually felt. This is consistent with Frith's (1989) hypothesis that individuals with autism have a preference for processing locally rather than globally. This tendency to process locally rather than globally has been demonstrated in a recent paper, where those with autism or Asperger syndrome attempted to justify a character's action by giving a locally rather than globally coherent inference (Jolliffe & Baron-Cohen, 1998a).

Although the clinical groups' difficulties suggest a weakness in processing mental state information in context, it is not entirely clear whether they could appreciate the mental states employed in this test. For example, it is not clear whether the participants could really appreciate what sarcasm means. Support for the notion that the clinical participants might have a deficient understanding of some of the mental state concepts employed comes from the within-group evidence, which found that the normal control participants performed better on the Mentalistic condition relative to their own performance on the Physical condition, whereas the

clinical groups showed no differential pattern of performance on their conditions. It was possible that the normal control group benefited from their understanding of different types of mental state, whereas the clinical groups did not. Alternatively, the within-group evidence might suggest something uniquely different about the contextual processing requirements of the two conditions. Further evidence for the clinical groups' difficulty on the Mentalistic condition being due to a lack of familiarity with the mental states employed comes from the disproportionate trouble they had with the concept of irony and double-bluff. Furthermore, the double-bluff stories needed to be understood at a third-order ToM level, since there is an extra level of embedding, that is, participants have to metarepresent "he *knows* they *think* he will *lie*", rather than "he *knows* he will *lie*."

Next we need to consider why the clinical groups not only perform less well on the Mentalistic condition in comparison to their normal control group, but also why they did not perform worse than their normal controls on the Physical condition, since this condition also requires the making of global inferences. Examining why the clinical groups perform normally on the Physical condition requires comparing the nature of the stimuli in this condition with that of the Mentalistic condition. The Mentalistic condition requires participants to infer the meaning of the utterance from the context provided, that is, they need to integrate the utterance with the context and thus have to use the context to extract meaning for the utterance. Whereas two of the six stimuli in the Physical condition also require participants to infer the meaning of the action from the context provided (i.e., they need to integrate the action with the context in the Army and Car stories), the remainder of the stories do not require such processing in context. Thus the remainder require a knowledge of why X-rays are normally taken, that the whites of eggs can be used for making meringues, that breaking security beams sets off alarms, and why glasses that correct long-sight are more likely to have been left at a Post Office than at other types of places such as a flower shop. These stories tap general knowledge more than processing in context. Also on these items more attention is perhaps drawn to the salient elements than occurs with the mentalistic stories. However, although the clinical groups did tend to make more errors on the two physical stories requiring the greater contextual processing and hence integration ability, the evidence for this type of processing being a problem for them was extremely weak. It seems that the stimuli tapping general knowledge tended to be performed slightly better, which may have assisted the clinical groups in performing at a normal level on the Physical condition. Finally, it is noteworthy that the crucial difference between the Physical and Mentalistic conditions is that a failure to take the context into account would lead to poor performance on the Mentalistic condition, but have a lesser effect on the Physical condition.

Although the main aim of this study was to attempt to replicate Happé's (1994) main finding of context-inappropriate explanations, a secondary aim was to see whether early language development could differentiate the two groups. As defined in our study, the individuals with Asperger syndrome do not exhibit a clinically significant delay in early language development, whereas the autism group recruited all had marked language delay. The results on the Strange Stories test suggests neither quantitative nor qualitative differences. The fact that there were no *significant* differences between groups suggests that the presence or absence of early language did not differentiate the two groups. However, although the clinical groups did not differ on any of the measures, it was noticeable that the autism group always performed at a level below that of the Asperger group (see Table II, III, and IV). Moreover, although 12 out of 17 of the participants with Asperger syndrome gave at least one context-inappropriate mental state answer, every one of the participants with autism gave at least one such answer. Whereas the results of the majority analysis suggest that this tendency to give context-inappropriate answers characterizes the majority of those on the autism spectrum, it is clear that this tendency was universal to the autism group but not the Asperger group. However, despite the autism group's relatively less efficient performance, the failure to find significant differences between the two clinical groups seems to suggest that on this test early language development does not discriminate between them. This is reminiscent of the findings from these same participants in other experiments assessing the processing of linguistic context for meaning (Jolliffe & Baron-Cohen, 1998a, in press). However, this finding is not specific to linguistic material, because recent visual experiments with these same participants also suggest that individuals with autism are relatively less efficient when it comes to noticing and identifying and an incongruent object within a scene (Jolliffe & Baron-Cohen, 1998b) and conceptually integrating multiple fragments of an object (Jolliffe & Baron-Cohen, 1998c). Thus the finding that the autism group was relatively but not significantly less efficient on the Strange Stories test might reflect more severe symptoms in childhood rather than language delay per se. However, this is certainly a fruitful line of enquiry for future research. Currently it seems that performance on the Strange Stories test seems to lend support to autism and Asperger syndrome being part of the same autistic continuum rather than being discrete conditions.

In conclusion, attempts to try to explain why the clinical groups performed poorly on the Mentalistic condition, but not on the Physical condition, in relation to controls, are unresolved. Thus we are still left with two possible reasons for their difficulties on the Strange Stories task. One is the requirement that the participant must infer the speaker's intended meaning not from the utterance but from the context in which it is embedded. The other is that there might have been a problem in appreciating some of the mental states employed in the Strange Stories test. Our conclusion is that although the Strange Stories test clearly identifies deficits in individuals on the autism spectrum, these are not "pure" deficits in that they could arise for theory of mind reasons, or central coherence reasons, or both. [For a more sensitive measure of adult ToM (and evidence of its impairment in these same patients) see Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997.] Because it is not entirely clear what the source of the clinical groups' difficulty is, future research needs to address this issue.

ACKNOWLEDGMENTS

We are grateful to a large number of people for helping us recruit participants. Lorna Wing, Wendy Phillips, the National Autistic Society, Pat Howlin, Clive Robinson, Ben Sacks, and Pam Yates. We are grateful to Ian Nimmo-Smith for advice on methodology and Francessca Happé for allowing us to use and adapt her materials. The first author was supported by an MRC Studentship during the period of this work. The Harold Hyam Wingate Foundation also provided her with valuable financial support. Both studies reported here were submitted in part fulfilment of her doctoral degree at the University of Cambridge.

APPENDIX

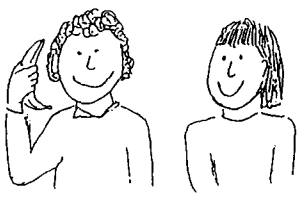
The Mentalistic Stories

Banana

Katie and Emma are playing in the house. Emma picks up a banana from the fruit bowl and holds it up to her ear. She says to Katie, "Look! This banana is a telephone!"

Is it true what Emma says?

Why does Emma say this?



Picnic

Sarah and Tom are going on a picnic. It is Tom's idea, he says it is going to be a lovely sunny day for a picnic. But just as they are unpacking the food, it starts to rain, and soon they are both soaked to the skin. Sarah is cross. She says, "Oh yes, a lovely day for a picnic alright!"

Is it true, what Sarah says?

Why does she say this?

The Physical Stories

Army

Two enemy powers have been at war for a very long time. Each army has won several battles, but now the outcome could go either way. The forces are equally matched. However, the Blue army is stronger than the Yellow army in foot soldiers and artillery. But the Yellow army is stronger than the Blue army in air power. On the day of the final battle, which will decide the outcome of the war, there is a heavy fog over the mountains where the fighting is about to occur. Low-lying clouds hang above the soldiers. By the end of the day the Blue army have won.

Q: Why did the Blue army win?

Glasses

Sarah is very long-sighted. She has only one pair of glasses, which she keeps losing. Today she has lost her glasses again and she needs to find them. She had them yesterday evening when she looked up the television programmes. She must have left them somewhere that she has been today. She asks Ted to find her glasses. She tells him that today she went to her regular early morning keep fit class, then to the post office, and last to the flower shop. Ted goes straight to the post office.

Q: Why is the post office the most likely place to look?

REFERENCES

American Psychiatric Association. (1987). *Diagnostic and statistical manual of mental disorders,* (3rd ed., Rev.) Washington, DC.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.) Washington, DC.

Baron-Cohen, S. (1989). The autistic child's theory of mind: A case of specific developmental delay. *Journal of Child Psychology and Psychiatry,* 30, 285–297.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press/Bradford Books.

Baron-Cohen, S. (1997). Hey! It was a joke! Understanding propositions and propositional attitudes by normally developing children, and children with autism. *Israel Journal of Psychiatry, 34,* 174–178.

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another advanced test of theory of mind: evidence from very high functioning adults with autism or Asperger syndrome. *Journal of Child Psychology and Psychiatry, 38,* 813–822.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a 'theory of mind'? *Cognition, 21,* 37–46.

Frith, U. (1989). *Autism: Explaining the enigma.* Oxford: Blackwell.

Happé, F. G. E. (1991). *Theory of mind and communication in autism.* Unpublished PhD thesis, University of London.

Happé, F. G. E. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition, 48,* 101–119.

Happé, F. G. E. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders, 24,* 129–154.

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66,* 843–855.

Howlin, P. (1995). *The Revised Howlin Screening Questionnaire,* St. George's Hospital Medical School, University of London.

Jolliffe, T., & Baron-Cohen, S. (1998a). Linguistic processing in high-functioning adults with autism or Asperger syndrome: Can global coherence be achieved? A further test of central coherence theory. Unpublished manuscript, University of Cambridge.

Jolliffe, T., & Baron-Cohen, S. (in press). A test of central coherence theory: Linguistic processing in high-functioning adults with autism or Asperger syndrome: Is local coherence impaired? *Cognition.*

Jolliffe, T. & Baron-Cohen, S. (1998b). A test of central coherence theory: Can adults with high-functioning autism or Asperger syndrome integrage objects in context? Unpublished manuscript. University of Cambridge.

Jolliffe, T. & Baron-Cohen, S. (1998c). A test of central coherence theory: Can adults with high-functioning autism or Asperger syndrome integrage fragments of an object? Unpublished manuscript. University of Cambridge.

Kanner, L. (1943). Autistic disturbances of affective contact. *Nervous Child, 2,* 217–250.

Klin, A., Volkmar, F. R., Sparrow, S. S., Cicchetti, D. V., & Rourke, B. P. (1995). Validity and neuropsychological characterisation of Asperger syndrome: Convergence with nonverbal learning disabilities syndrome. *Journal of Child Psychology and Psychiatry, 36,* 1127–1140.

Miller, J. N., & Ozonoff, S. (1996). Did Asperger's cases have Asperger disorder? A research note. *Journal of Child Psychology and Psychiatry, 38,* 247–251.

Minshew, N. J., Goldstein, G., Muenz, L. R., & Payton, J. B. (1992). Neuropsychological functioning in non-mentally retarded autistic individuals. *Journal of Clinical and Experimental Neuropsychology, 14,* 749–761.

Ozonoff, S., & Miller, J. N. (1996). An exploration of right-hemisphere contributions to the pragmatic impairments of autism. *Brain and Language, 52,* 411–434.

Perner, J., Frith, U., Leslie, A. M. & Leekam, S. R. (1989). Exploration of the autistic child's theory of mind: Knowledge, belief and communication. *Child Development, 60,* 689–700.

Rumsey, J. & Hamburger, S. (1998). Neuropsychological findings in high-functioning men with infantile autism, residual state. *Journal of Clinical and Experimental Neuropsychology, 10,* 201–221.

Rumsey, J. M. & Hanahan, A. P. (1990). Getting it "right": Performance of high-functioning autistic adults on a right-hemisphere battery. *Journal of Clinical and Experimental Neuropsychology, 12,* 81.

Tantam, D. (1992). Characterizing the fundamental social handicap in autism. *Acta Paedopsychiatrica, 55,* 88–91.

Terman, L. M. & Merrill, M. A. (1973). *Stanford-Binet Intelligence Scale Form L-M.* (3rd edition). Boston: Houghton Mifflin.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale - Revised.* New York: Psychological Corp.

Wing, L. (1981). Asperger syndrome: A clinical account. *Psychological Medicine, 11,* 115–129.

World Health Organization. (1994). *International classification of diseases and related health problems* (10th ed.) Geneva.