Journal of Autism and Developmental Disorders, Vol. 24, No. 2, 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[1]

Francesca G. E. Happé[2]
Medical Research Council Cognitive Development Unit

Research has suggested that the core handicaps of autism result from a specific impairment in theory of mind (ToM). However, this account has been challenged by the finding that a minority of autistic subjects pass 1st- and even 2nd-order ToM tests while remaining socially handicapped. In the present study, able autistic subjects who failed ToM tasks, those who passed 1st-order, and those who passed 2nd-order tasks were tested with a battery of more naturalistic and complex stories. Autistic subjects were impaired at providing context-appropriate mental state explanations for the story characters' nonliteral utterances, compared to normal and mentally handicapped controls. Performance on the stories was closely related to performance on standard ToM tasks, but even those autistic subjects who passed all ToM tests showed impairments on the more naturalistic story materials relative to normal adult controls.

INTRODUCTION

The "theory of mind" deficit account of autism suggests that the communication, socialization and imagination handicaps of autistic individuals

[1]This work was completed while the author held an MRC studentship, and was submitted in partial fulfillment of a Ph.D. degree, University of London. I am grateful to the teachers and pupils of the following schools, for their kind assistance: Broomhayes School, Heathlands School, Helen Allison School, Sybil Elgar School, Springhallow School, Ramsden School. I also thank Fran Siddons for her help in collecting the data, and Uta Frith and all my colleagues at the CDU for invaluable discussions.
[2]Address all correspondence to Francesca Happé, MRC Cognitive Development Unit, 4 Taviton Street, London WC1H 0BT, England.

0162-3257/94/0400-0129$07.00/0

spring from their inability to represent and attribute mental states (Frith, 1989; Leslie, 1987, 1988). Baron-Cohen, Leslie, and Frith (1985) showed that 80% of autistic subjects failed to attribute a false belief to a character in an acted out story. In contrast, 80% of mentally handicapped and normal 4-year-old subjects were able to predict the character's behavior on the basis of their exposure history and resulting (inferred) false belief. Subsequent studies have replicated the failure of a majority of autistic subjects to attribute a false belief (e.g., Sodian & Frith, 1992), and although there are alternatives to the theory of mind account (e.g., Hobson, 1989, 1990), it has been successful at the very least in making concrete predictions about assets and deficits in autism (Frith, 1989; Firth, Morton, & Leslie, 1991).

The finding that only approximately 20% of autistic subjects pass a first-order false belief task has been taken as strong support for an explanation of the autistic handicap in terms of a lack of theory of mind (ToM). However, that even 20% should pass has been seen by some critics (e.g., Bowler, 1992; Ozonoff, Rogers, & Pennington, 1991) as damaging to the explanatory power of the theory. Indeed, these authors have found that in groups of autistic subjects selected for normal verbal IQ, the success rate on theory of mind tasks is much higher, with autistic subjects performing indistinguishably from controls on even second-order tasks. The response to such criticism has taken two directions (discussed further in Happé, 1993). The success of these subjects could be regarded as genuine proof of their possessing a theory of mind—in which case their handicaps may be seen as due to a gross delay in acquisition (Baron-Cohen, 1989) or to an additional, remaining cognitive impairment (Happé, in preparation b). Alternatively, their success could be seen not as proof of theory of mind ability but rather as evidence of the "hacking out" of some strategy for solving the tasks (Frith et al., 1991). If autistic subjects who pass theory of mind tasks succeed using a non-ToM strategy, this would explain why, despite perfect test performance, these subjects are still socially handicapped. The strategy such subjects could be using might be flexible enough to be applied successfully to slightly different surface forms of the same task, or might succeed only if the elements of visual access and information are spelled out (as they are in the false belief and deception tasks, but not in life).

To test these possibilities, the present study presented autistic subjects with a set of vignettes or stories about everyday situations where people say things they do not mean literally. It was hoped that these stories would present a somewhat more naturalistic challenge to the subjects than did the acted out ToM battery tasks. The aim, then, was to extend the range of tasks involving theory of mind to a more contextually embedded and

realistic form, which might be expected to trip up even those subjects who succeeded on the previous, simplified tasks.

The stories were not imaginative or highly fictional. They were simple accounts of events, which concerned the different motivations that can lie behind everyday utterances that are not literally true. So, for example, if someone asks your opinion of a new dress which you actually think is hideous, you might say it was nice for a variety of different reasons; to spare their feelings, to mislead them into wearing it and looking awful, to be sarcastic, or to be funny. In everyday life these different motivations are distinguished by many factors, such as preceding context, emotional expression, and relationship between speaker and hearer. The stories used here were written to be largely unambiguous, so that only one interpretation of the situation would be made by normal and nonautistic mentally handicapped subjects. The prediction was that autistic subjects would have greater difficulty with the stories than would the controls, but that autistic subjects' performance would show a strong relation to their performance on the theory of mind battery. The precise relation of the subjects' performance on the two test batteries should reveal something about the cognitive processes underlying their success or failure on the traditional (false belief) tests of attribution of mental states. Control stories, involving only the understanding of physical events, were also given to the subjects, to check the generality of any comprehension deficit which might emerge regardless of story content.

METHOD

Subjects

The stories were given to 24 autistic subjects. This group included 18 subjects who had previously completed the theory of mind battery (see Happé, in preparation a). All of these subjects passed first-order false belief (FB1) tasks, and this group is referred to below as the "able autistics." In addition, 6 autistic subjects who failed to pass first-order false belief tasks, or any subsequent tasks from the battery, received the story materials. These subjects are referred to as the "no theory of mind" group ("no-ToM" autistic group).

To examine the nature of the able autistic subjects' success, subjects were selected from the 18 individuals to form two groups. Subjects who performed most consistently well at first-order tasks but failed second-order tasks made up a "first-order theory of mind" group ("1st-order" autistic group), and subjects who performed consistently well at both first- and sec-

ond-order tasks made up a "second-order theory of mind" group ("2nd-order" autistic group). Other subjects, who performed inconsistently (or were unavailable for later testing) were excluded from these groups. It is important to stress that the labels given to the two theory of mind groups of autistics were not intended to make strong claims about their underlying theory of mind competence. Rather the labels refer to the subjects' level of *performance* on the theory of mind tasks. The nature of the processes underlying their successful performance was still an open question at this point. The subjects in the 1st-order and 2nd-order groups were, however, considered the best candidates for possessing "true" theory of mind, since their performance was consistent at one or both of the two levels.

Controls for this experiment were (a) 11 subjects with mental handicap (MH) who had also taken part in the theory of mind battery study (all of whom passed first- and second-order tasks), plus 2 MH adults who did not score perfectly on first-order false belief tasks but who performed well otherwise across the battery; (b) 26 normal children who had taken part in the ToM battery (all of whom passed first- and second-order tasks); and (c) a group of 10 normal adults. The adults were volunteers; three were in their final school years, and four were students at London University. This control group was included to establish that subjects presented with simple stories would not make errors due to trying to out think the experimenter—which might be a possible explanation of poor performance in the able autistic subjects. Subject characteristics can be found in Table I. Both the able autistic group and the MH control group showed a preponderance of males, whereas the normal adult and child groups were made up of equal numbers of male and female subjects.

As can be seen in Table I, the autistic subjects who failed the theory of mind tasks had significantly lower VIQs than those who passed. Since the verbal subtests of the WISC/WAIS-R make pragmatic demands upon the subject, it may be that theory of mind task performance and verbal IQ (insofar as it inadvertently measures communicative competence) are not independent measures (Happé, 1991; Happé, 1993). For this reason, VIQ differences were not covaried out in analyzing the results, since this would appear to overcontrol for the variable of interest (theory of mind ability). It is important to note, however, that the no-theory-of-mind autistics were approximately matched for VIQ with the MH controls (although tested on different verbal assessments), and so any relative deficit in the autistic group cannot be due simply to generally low verbal ability. Similarly, the 1st-order and 2nd-order ToM groups did not differ significantly in VIQ, and so differences between these two groups are also not simply due to general ability differences.

Table I. Subject Characteristics for "Strange Stories" Battery

Subject	*n*	Age *M*	Age Range	VIQ *M*	VIQ Range
Young normal control	26	8.6	6.6–9.7	—	—
Normal adult	10	20.5	15–24	—	—
MH control	13	19.4	12–38	56.9[a]	40–89
Able autistic (FB1 = 4/4)	18	20.6	8.9–45.1	87.3[b]	64–101

Autistic subjects grouped by theory of mind (ToM) performance

			Age		VIQ(WISC/WAIS)	
		M:F	*M*	Range	*M*	Range
No-ToM autistic	6	2:1	17.6	13.5–28.2	62.3[c,d]	52–76
1st-order ToM autistic	6	5:1	16.7	8.9–24.6	81.8[c]	65–100
2nd-order ToM autistic	6	6:0	17.7	11.5–25.5	95.8[d,e]	90–101

[a]From BPVS.
[b]From WISC-R or WAIS. ANOVA $F(2, 14) = 11.39$, $p < .001$.
[c]No-ToM autistic < 1st-order ToM autistic subjects, Tukey's test $p < .05$.
[d]No-ToM autistic < 2nd-order ToM autistic subjects, Tukey's test $p < .01$.
[e]From 5 subjects only.

Materials

The theory of mind battery used to group the autistic subjects consisted of four 1st-order false belief tasks (FB1) such as the "Sally-Ann task" used by Baron-Cohen et al. (1985), and Leslie and Frith (1988), and the "smarties task" of Perner, Frith, Leslie, and Leekam (1989); two 1st-order deception tasks modeled on Wimmer and Perner's (1983) tasks; two 2nd-order false belief tasks (FB2) as used by Baron-Cohen (1989); and two 2nd-order deception tasks devised by the author, involving a double-bluff situation.

The set of "Strange Stories" consisted of 24 short vignettes, each accompanied by a picture and two test questions; the comprehension question "Was it true, what X said?," and the justification question "Why did X say that?" There were 12 types of story, and two examples of each story type. The 12 story-types comprised Lie, White Lie, Joke, Pretend, Misunderstanding, Persuade, Appearance/Reality, Figure of Speech, Sarcasm, Forget, Double Bluff, and Contrary Emotions. A set of six control "physical stories" were also given to the subjects. These stories did not involve mental states, but instead described an unforseen outcome with a mechanical-physical cause (e.g., a power-cut causing a meal to be undercooked). Examples of the experimental stories and control stories can be found in the Appendix.

Procedure

The MH and autistic subjects were tested alone, in a quiet room in the subject's school or home. The set of stories was introduced as follows; "Here are some stories, and some questions. I'm going to read out the stories and I'd like you to listen carefully, and help me with the questions at the end of each story." Most subjects finished all the stories in one testing session, but breaks were given as needed. Each story was read out to the subject—except where the subject preferred to read the story out loud to the experimenter. The story remained in front of the subject throughout to minimize memory requirements. At the end of the story the subject was asked the two (or sometimes three) test questions. The first question, "Was it true, what X said?" was treated as a test of comprehension. Therefore, although the first response was recorded, if the answer was wrong the story was read out again, until the subject answered correctly or justified their answer and appeared to understand (e.g., "it's not literally true, but it's not a lie"). The second question, "Why did X say that?" was then asked, and the subject's answer was recorded in full on scoring sheets, for later analysis. Positive comments were made throughout the testing session to encourage the subject, but no feedback was given about the correctness of the answers. Administration was adjusted to the requirements of the subjects, with repetition where necessary, and so length of time required varied greatly (from approximately 20 minutes to 1 hour). Prompts were given only in order to establish sufficient understanding of each story to correctly answer the "Was it true?" question. So, for example, a subject might be reminded in the white lie story that Peter did not like his aunt's hat, by the additional prompt, "What did Peter think of the new hat?" In a few cases, subjects would not answer the "Why" question. Details of the frequency of such omissions are given below.

The normal subjects received the stories in a slightly different way. The normal adults were given the set of stories to take away and complete overnight with specific instructions not to ask advice from other people. The young normal children were all from a village primary school in Sussex. The stories were given to their class teacher who incorporated them into the day's activities. The children spent 10 minutes each day working through the stories individually and on their own. The teacher insured that children did not confer, and she herself only helped the children with problems in reading the stories, understanding specific words, or spelling their answers.

For each subject the stories were presented in randomized order, and the two examples of each story type were not placed together.

Scoring Procedure

The justifications given in response to the "Why" question were rated as either correct or incorrect. A justification could be incorrect because it involved errors about the facts given in the story, or because it involved an inference that was inappropriate as a reason for the story character's utterance. For example, in the lie story, in which Anna breaks a vase but tells her mother that the dog did it, the justification "Anna did not break the vase," would be scored as incorrect because it includes a factual error. A justification that "Anna was just joking" would also be scored as incorrect, because in the context of the story it is not appropriate to interpret her utterance as a joke. This latter judgment of appropriateness was clearly a subjective one, but a good degree of interrater agreement was reached as to the correctness/incorrectness of the subjects' answers (see below).

The justifications were also scored as either involving mental states or physical states. Many of the story characters' utterances could be justified correctly either in terms of mental states or physical states. For example, in the joke story where a boy calls a dog an elephant, this can be correctly explained by the physical justification, "the dog is big like an elephant," or the mental state justification, "He's just joking." Mental state answers included all those that referred to thoughts, feelings, desires, traits, and dispositions. Mental state justifications included terms such as like, want, happy, cross, afraid, know, think, joke, pretend, lie, to fool someone, expecting. Justifications were scored as physical state when they referred to nonmental events—physical appearance, action of objects, physical events, and outcomes. Physical state answers included terms such as big, looks like, is shaped like, to get rid of them, to sell them, because of the X (object), to not get X (physical outcome, e.g., put in jail, have a filling).

In each case only one score was given per story, giving the subject credit for their "best" answer. That is, if a subject gave one correct answer and one inappropriate answer, the correct answer was taken. Similarly, if a subject's answer appealed to both physical and mental states, the justification would be scored as mental state.

The subjective judgment of justifications made covalidation of the scoring necessary to establish validity. The justifications given by every subject tested, to one story of each type, were given to a second rater, who was naive to the hypothesis being tested, and blind to the identity and diagnosis of the subjects. Examples of the types of justification, as given to the corater, can be seen in the Appendix. The degree of concordance was calculated for each story-type separately, and ranged from 92 to 100%. In all, 22 justifications received discordant ratings; 9 of these were given by autistic subjects, 9 by MH subjects, and 4 by young normal subjects. Although the

normal adults and children were less likely than the other groups to provide answers over which the raters disagreed (in general providing clearly correct and common justifications), the autistic and MH groups did not differ from one another in the number of hard-to-rate answers given.

Omissions were recorded on one or more stories for 2 of the 10 adult controls, 14/26 normal children, 4/13 MH subjects, and 12/24 autistic subjects. The number of autistic subjects failing to give an answer to at least one item did not differ significantly from the number of MH, $\chi^2(1) = 1.24$, young normal children, $\chi^2 = 0.08$, or adult controls, $\chi^2 = 2.58$, who scored at least one omission. However, the average number of omissions per subject was greater in the autistic and young normal groups than in the MH and normal adult groups. Omissions occurred for 6.6% of all items in the autistic group, and to 4% of items in the young normal controls, compared with 1.9 and 1.3% for the MH and adult controls, respectively. The higher incidence of omissions by the autistic subjects (between 0 and 9, average 1.6) may be due to the relatively greater difficulty of the test questions for this group—the autistic subjects in the 2nd-order ToM group gave on average only 0.2 omissions whereas the other autistic subjects gave on average 2.1 omissions. The relatively high incidence of omissions in the young normal sample (number of omissions by a subject ranged from 0 to 5, with an average of 1) probably reflected the rather different testing procedure in this group. The difference in number of omissions across the different diagnostic groups, does not invalidate the analysis of the results below. Results are analyzed in terms of total correct, a score to which omissions do not contribute, reflecting their status as probable indications of inability to answer the test question. Results in terms of number of correct mental state answers are also analyzed. The greater number of omissions by the autistic subjects is unlikely to contribute to differences (fewer correct responses are predicted from the autistic vs. control groups) here, since the autistic groups do not give significantly fewer responses of any other type (mental state incorrect, and right or wrong physical state). Analysis of mental state errors will, presumably, be underestimated rather than overestimated by the tendency of autistic subjects to give fewer responses (and so makes for a conservative test of the prediction that autistic subjects will have selective problems with mental state understanding).

RESULTS

The physical stories proved to be very easy for subjects in all groups. No subject scored less than 5 out of 6, and no group differences emerged. Since all groups were at ceiling, performance on these stories cannot be

contrasted meaningfully with performance on the Strange Stories. Therefore the physical stories can only be treated as a screening device. However, these control stories are useful in demonstrating that where mental states are not involved, autistic subjects can understand and answer questions about simple stories.

On the Strange Stories all but 10 subjects scored 24 out of 24 on their first attempt at the "Is it true?" question, which was regarded as a test of linguistic comprehension. The groups did not differ significantly on this measure; 3 of the 24 autistics, 2 of the 13 MH controls, 4 of the 26 normal children, and 1 normal adult scored 22 or 23 out of 24. These errors were mainly due to confusion over the sense of "true" intended in the question (e.g., some normal subjects said that although it was not literally true that the girl in the figure of speech story had a frog in her throat, it was not untrue to use that expression).

The justifications given by the subjects in response to the "Why" question were scored as correct or incorrect, and as concerning mental or physical states. The results for the able autistic group and controls can be seen in Table II. As can be seen, the able autistic group, who were matched with the controls in passing first-order false belief tasks, showed significantly worse performance on the strange stories than the control groups. Their significantly lower total score, and their tendency to give incorrect rather than correct mental state justifications, is striking—especially in view of their verbal IQ, which was higher than the MH controls', and their age and experience, which was greater than that of the young normal controls.

The mean total number of mental state justifications (max = 24) given by the able autistic group was 15.9 (range 8–23) versus 16.5 by the MH group (12–22) and 18.2 by the young normals (15–22). Thus the autistic group did not differ in the tendency to use mental state language in answering the Strange Stories "Why" question, but rather differed only in the appropriateness or accuracy of the mental states attributed.

The results also show that where there is an acceptable physical state justification (e.g., for saying a banana is a phone: "Because it looks like one") the autistic subjects perform as well as the young normals and MH controls. All three groups differ from the normal adults, who tend not to give physical state justifications, referring instead to mental states (in the banana example: "She's just pretending its a phone").

Results for Autistic Subjects by Theory of Mind Group

The Strange Stories were also intended to explore the validity of the ToM battery tests and the successful performance of some of the autistic

Table II. Number of Justifications

Subject group	*n*	Total score (Max = 24)	Correct justifications		Incorrect justifications	
			Mental	Physical	Mental	Physical
Normal adult	10					
M		23.7	22.5[b]	1.2[c]	0.0	0.0
SD		0.7	2.1	1.2		
Range		22–24	18–24	0–5		
Normal children	26					
M		21.0	16.7	4.4	1.9	0.2[e]
SD		3.0	3.1	1.9	2.0	0.4
Range		11–24	10–22	1–9	0–8	0–1
MH control	13					
M		21.4	15.9	5.5	0.7	1.6[e]
SD		2.0	2.3	2.2	0.6	1.4
Range		17–24	12–20	2–10	0–2	0–4
Able autistic	18					
M		15.7[a]	11.1[b]	4.6	4.8[d]	1.7[e]
SD		4.0	4.5	2.4	2.8	1.9
Range		9–21	3–19	0–9	0–11	0–5

[a]A one-way ANOVA showed a significant effect of diagnostic group on total score, $F(3, 63) = 20.10$, $p < .000$, and Tukey's test showed that the able autistic subjects were significantly worse than all other groups ($p < .01$).

[b]ANOVA, $F(3, 63) = 26.8$, $p < .000$; Tukey's test shows that able autistic subjects gave significantly fewer correct mental state justifications than all other groups ($p < .01$), and normal adults gave significantly more than normal children or MH controls ($p < .01$).

[c]ANOVA, $F(3, 63) = 8.5$, $p < .000$; Tukey's test shows that normal adults gave significantly fewer correct physical state justifications than all other groups ($p < .01$).

[d]ANOVA, $F(3, 63) = 18.4$, $p < .0000$; Tukey's test shows that autistic subjects gave more incorrect mental state justifications than any other group ($p < .01$).

[e]ANOVA, $F(3, 63) = 18.4$, $p < .000$; Tukey's test shows that normal children used fewer incorrect physical state justifications than did MH or autistic subjects ($p < .01$), normal adults used fewer incorrect physical state justifications than either MH ($p < .05$) or autistic subjects ($p < .01$).

subjects. The results for the autistic subjects grouped by their performance on the ToM battery can be seen in Table III. As can be seen from the ranges in Table III, there was little overlap in the total number of correct justifications between the three ToM groups. The three groups differed significantly from each other in total score, with autistic subjects in the no-ToM group scoring least and those in the 2nd-order ToM group scoring most. This supports the validity of the ToM battery tasks, suggesting that differences in performance on that battery reveal real underlying differences in the ability to attribute mental states correctly in a variety of tasks.

The three groups of autistic subjects were also compared individually with the control groups with one-way ANOVA followed by Tukey's tests. The no-ToM group of autistic subjects was significantly worse than all three control groups: The total correct for this group was significantly less than

Table III. Number of Justifications; Autistic Subjects Grouped by ToM Performance

Subject group	Total score (Max = 24)	Correct justifications		Incorrect justifications	
		Mental	Physical	Mental	Physical
No-ToM autistic					
M	7.5[a]	5.0	2.5	11.0[c]	4.7[d]
SD	1.1	1.5	1.1	4.2	2.9
Range	6–9	3–7	1–4	8–16	2–9
1st-order autistic					
M	12.8[a]	7.7	5.1	6.0	3.0
SD	2.4	3.7	3.1	2.2	2.4
Range	9–16	3–13	0–9	4–10	0–6
2nd-order autistic					
M	20.0[a]	15.3[b]	4.7	3.3	0.5[d]
SD	1.5	3.1	2.4	3.3	0.5
Range	17–21	11–19	1–7	1–6	0–1

[a]ANOVA, $F(2, 14) = 66.7, p < .000$; Tukey's test shows that the total score of no-ToM group is significantly less than that of other two groups ($p < .01$), and the 2nd-order group scored significantly higher than 1st-order or no-ToM groups ($p < .01$).
[b]ANOVA, $F(2, 14) = 19.2, p < .000$; Tukey's test shows that the 2nd-order ToM autistic subjects gave more correct mental state justifications than the other two groups ($p < .01$).
[c]ANOVA, $F(2, 14) = 10.6, p < .002$; Tukey's test shows that the no-ToM autistic subjects gave more incorrect mental state justifications than did the 1st-order ToM group ($p < .05$) or 2nd-order ToM group ($p < .01$).
[d]ANOVA, $F(2, 14) = 5.6, p < .016$; Tukey's test shows that the no-ToM autistic subjects gave more incorrect physical state justifications than the 2nd-order ToM group ($p < .05$).

the MH, young normal, or adult controls, $F(3, 50) = 60.6, p < .000$; Tukey $p < .01$ for each pairwise comparison. Similarly the no-ToM group gave significantly fewer correct mental state justifications, $F(3, 50) = 51.5, p < .000$; Tukey $p < .01$, and more incorrect mental state justifications than any of the control groups, $F(3, 50) = 44.9, p < .000$; Tukey $p < .01$. The 1st-order ToM group of autistic subjects also performed poorly compared with the controls. Their total correct was significantly lower than any of the three control groups, $F(3, 51) = 26.3, p < .000$; Tukey $p < .01$ for all pairwise comparisons. They also gave fewer correct, $F(3, 51) = 34.9, p < .000$; Tukey $p < .01$, and more incorrect mental state justifications than the controls, $F(3, 51) = 20.5, p < .000$; Tukey $p < .01$.

The autistic group most interesting to compare with the controls was that consisting of subjects who passed the 2nd-order ToM tasks in the previous experiment. It had been hoped that the Strange Stories might be more naturalistic than the usual theory-of-mind tasks, and so might reveal the true and complex handicaps that even this group of subjects appeared to display in everyday life. The results showed some evidence of this. The 2nd-order group had a significantly lower total correct than the normal

adult controls, $F(3, 51) = 4.1, p < .01$; Tukey $p < .05$. The normal adults also gave significantly more correct mental state justifications than the 2nd-order ToM autistic subjects, $F(3, 51) = 14.6, p < .000$; Tukey $p < .01$, and than the MH and young normal controls ($p < .01$). Last, these autistic subjects gave significantly more incorrect mental state answers than the normal adults and even than the (much lower IQ) MH controls, $F(3, 51) = 7.9, p < .000$; Tukey $p < .01$. The significance of these results is discussed below, where the case is made that for these very able autistic subjects normal adults form the most appropriate control group.

The results were also analyzed in terms of frequencies. The number of subjects in each group giving incorrect mental state justifications was compared, since this type of answer had proved a discriminative measure in the foregoing analyses. In addition, the battery of stories was decomposed into the 12 story-types for this frequency analysis, allowing an exploration of the relative difficulty of the different scenarios described. The frequencies expressed as numbers per group and as percentages can be seen in Table IV. The normal adults do not appear in Table IV, since no subject in this group gave an incorrect mental state justification. However, significant differences shown between autistic and young normal and MH control groups also hold between autistics and normal adults. Table IV shows that more able autistic subjects made mental state errors than any other group—and this was true for 10 of the 12 story-types. Only on stories

Table IV. Subjects Giving at Least One Incorrect Mental State Justification

	MH controls (n = 13)		Normal children (n = 26)		Able autistic subjects (n = 18)	
Story type	n	%	n	%	n	%
Pretend	1	7.7	2	7.7	8	44.4[a]
Joke	0		3	11.5[b]	5	27.8[a]
Lie	0		0		4	22.2[a]
White lie	0		4	15.4	9	50.0[a]
Figure of speech	1	7.7	8	30.8	5	27.8
Misunderstanding	1	7.7	2	7.7	6	33.3[a]
Double bluff	3	23.1	2	7.7	8	44.4[a]
Sarcasm	2	15.4	10	38.5[b]	10	55.6[a]
Persuasion	0		4	15.4[b]	7	38.9[a]
Contrary emotion	1	7.7	6	23.1[b]	1	5.6
Appearance/reality	0		0		5	27.8[a]
Forget	0		3	11.5[b]	5	27.8[a]

[a]Significantly more autistic subjects than young normal and MH controls make mental state errors, on chi-square test and Page's trend test $p < .05$ or $p < .01$.

[b]Significantly more of the young normals than the MH controls make mental state errors, on chi-square and Page's trend test $p < .05$ or $p < .01$.

about figures of speech and about contrary emotions did autistic subjects make as few errors as controls. This is interesting in view of the fact that figures of speech may be learned without any appreciation of mental states, since they are frozen expressions. It is also intriguing that the autistic subjects were not troubled by the contrary emotion stories.

The numbers of autistic subjects, grouped by performance on the ToM battery, who gave at least one incorrect mental state justification can be seen in Table V below. The frequencies are also given as percentages to give an impression of how these data compare with those from the control groups (given in Table IV). Table V shows that three of the story-types discriminated well between the three sets of autistic subjects grouped by ToM performance. On stories concerning joking, lying, and persuasion more of the subjects in the no-ToM group made mental state errors than did subjects in the 1st-order group, who in turn were more likely to make such errors than the 2nd-order autistic subjects.

DISCUSSION

The battery of Strange Stories proved useful in discriminating between the control subjects and those autistic subjects who failed all ToM

Table V. Subjects Giving at Least One Incorrect Mental State Justification

	No-ToM autistic subjects (n = 6)		1st-order autistic subjects (n = 6)		2nd-order autistic subjects (n = 6)	
Story type	n	%	n	%	n	%
Pretend	4	66.7	3	50	1	16.7
Joke	4	66.7[a]	2	33.3	1	16.7[b]
Lie	5	83.3[a]	3	50	0[b]	
White lie	6	100	5	83.3	1	16.7[b]
Figure of speech	3	50	3	50	1	16.7
Misunderstanding	5	83.3	2	33.3	2	33.3
Double bluff	3	50	3	50	2	33.3
Sarcasm	5	83.3	3	50	3	50
Persuasion	6	100[a]	4	66.7	2	33.3[b]
Contrary emotion	3	50	1	16.7	0	
Appearance/reality	0		1	16.7	2	33.3
Forget	2	33.3	2	33.3	2	33.3

[a]On chi-sqare and Page's trend test, significantly more subjects in the no-ToM group than in the other two groups made mental state errors, $p < .05$ or $p < .01$.

[b]On chi-square and Page's trend test, significantly fewer subjects in the 2nd-order ToM group than in the other two groups made mental state errors, $p < .05$ or $p < .01$.

tasks. Only one subject from the control groups scored a total correct of less than 17 out of 24 on the stories (this subject was a young normal child who had just turned 8, who scored 11/24). In contrast, autistic subjects in the no-ToM group scored no more than 9/24, and none of the subjects in the 1st-order autistic group scored 17 or more. Total score on the stories also discriminated well between the three groups of autistic subjects, allocated to groups by their performance on the theory of mind battery. The almost complete lack of overlap between the groups on this measure suggests that the varied and complex materials of the Strange Stories battery tap the same underlying ability as the ToM battery, despite their quite different format. This result therefore supports the validity of the theory of mind battery tasks, since understanding of false belief and deception there predicted understanding of such diverse but conceptually connected situations as pretence, joking, persuasion, telling a white lie, and so forth, in the Strange Stories.

While it is possible that the subjects in the no-ToM group performed worse than those who passed either 1st- or 2nd-order ToM tasks simply due to their lower verbal IQ, this explanation appears unlikely in view of the good performance of the MH controls. The MH subjects, who match the no-ToM group on verbal IQ, performed significantly better than this group on the Strange Stories (total score: MH controls *M* 21.4, range 17–24, vs. *M* 7.5, range 6–9). Thus high verbal IQ cannot be essential for this task. This suggests that the group differences found may be due to real underlying differences in understanding of mental states, that is, theory of mind.

Groups were also well distinguished in terms of number of correct mental state justifications given. Here the control subjects never gave fewer than 12 correct mental state answers (with one exception: the same young normal, who gave only 10 correct mental state answers), while the no-ToM group of autistic subjects never gave more than 7, and only one of the autistic subjects in the 1st-order group gave more than 10 (one subject, 13) such answers. By contrast the autistic subjects who passed 2nd-order ToM tasks also performed well on the stories, giving at least 11 correct mental state answers.

It is not clear why the joking, lying, and persuasion stories in particular should show significant differences between the three groups of autistic subjects. It is probably unwise to draw strong conclusions from such small numbers, but it may be that these stories are at a level of difficulty that best reveals real underlying differences in the three groups' ability to attribute mental states. In contrast, stories about sarcasm, double bluff, and so on may be too difficult, and stories about appearance/reality, forgetting, and so forth too easy.

In the no-theory-of-mind group there was a tendency to give one mental state justification again and again, interspersed with physical state answers. For example, one subject gave the same response—"she/he's having a joke"—for 15 of the 24 stories. While this response was correct for the joke stories it was highly inappropriate for the lie, white lie, misunderstanding, persuasion, and forgetting stories for which it was also given. One 18-year-old boy with a full-scale IQ of 85 prefixed 14 of his 24 answers with the verbs "to think" or "know," but evidenced little understanding of the mental states involved; for example, "he thinks a lawn mower cut her hair" for the joke story, "she doesn't know he doesn't keep pigs in his room" for the figure of speech story, and "she thought it was a rabbit, and she didn't know she didn't want the book" for the white lie story. A third subject used the answers "she/he made a mistake" and "she/he couldn't make up her/his mind" repetitively and largely inappropriately in response to the stories. This "parroted" mental state language may account for the acceptable mental state answers (*M* number 5) credited to this group. The tendency to use a particular mental state justification throughout did not appear to do with conditioning in the test situation; it was not the case, for example, that the explanation used was the correct answer to the first story read. In any case, the experimenter's response was always positive, and no differential feedback was given for right versus wrong answers. None of the MH controls showed this pattern of perseverative responding, and so it is unlikely to represent simply an attempt to generate answers where low ability makes material hard to comprehend. It therefore seems that these subjects came to the test situation equipped with one or two explanations for why people say puzzling things. It is possible that these explanations had been told to the subjects in response to questions about particular situations, and that the subjects noted them without understanding the precise nature of the context in which they apply. Frith (1989) and Happé (in preparation b) discuss why autistic subjects may find it hard to perceive such a context.

On the measures discussed so far the 2nd-order ToM group of autistic subjects performed indistinguishably from the young normal and MH controls. It should be remembered, however, that this group of autistic subjects had a mean VIQ of 95 compared to the MH group's mean VIQ of 56. There was in fact no overlap in the VIQs of these two groups, and one might conclude that since the autistic subjects in this group were all of approximately normal IQ (and only one was under 16 years old) the appropriate control group would be a group of normal adults. Comparison with the normal adults tested here reveals that the Strange Stories did in fact expose deficits in even this very able group of autistic subjects. The autistic subjects gave significantly fewer correct mental state answers and

made significantly more errors in attributing mental states. While every one of the 2nd-order group gave at least one (and on average 3) incorrect mental state attribution, not one of the normal adults made such a mistake. The errors made by the autistic subjects in the 2nd-order group were striking; one subject said that the mother in the sarcasm story said what she said ("That's what I call politeness!") "not to shock her daughter"; another subject explained the utterance in the pretence story ("This banana is a telephone") by saying "she said it to fool her"; a third subject's justification for the double-bluff soldier story was "he just wanted to tell the truth," and for the persuasion story he described the threat to drown the kittens as "just a joke." These results suggest that the battery of Strange Stories may be a more sensitive and naturalistic test of theory of mind ability than the relatively artificial and simplified false belief and deception tasks of the theory of mind battery. The results on the stories may more closely reflect the real life difficulties in understanding other minds that even the most able autistic subjects seemed to have. The comparison with the normal adults in this experiment is, however, imperfect, since these adults were students and probably of significantly higher IQ than even the 2nd-order ToM autistic subjects. It is therefore important to compare such able autistic subjects with less intelligent normal adults.

The tendency to give *correct* physical state answers (e.g., "because the dog is big like an elephant," "because the banana is shaped like a phone") to the "Why" question appears to be related to mental age in the nonautistic groups. While the normal adults rarely gave such answers (4 of the 10 adults gave between 1 and 5 such answers), they were a regular feature of the performance of the young normal children and the MH controls. In both these groups every subject gave at least one correct physical state justification, numbers varying between 1 and 10. Most of the autistic subjects also used such justifications, although here this appeared to have less to do with ability, and even the most able autistic subjects (who had IQs in the normal range and were as old as the normal adults) all gave at least one such answer.

The results for *incorrect* physical state answers are harder to interpret, since this category includes both answers that are factually incorrect about physical aspects of the stories (e.g., "the dog broke the vase") and answers that refer inappropriately to physical events where mental events are the appropriate focus for the justification (e.g., "the vase is broken"). This makes it hard to know whether a large number of incorrect physical state justifications reveals a subject's general inability to understand the story events, or their specific difficulty with mental state attribution (leading them to appeal to physical states as explanations for what is said). The results show that normal children and adults seldom give incorrect physical

state justifications, whereas autistic and MH subjects make significantly more and do not differ from one another in this respect.

Very different processes may, however, underlie the similar scores on this measure in the MH and autistic groups. In the 1st-order ToM group many of the autistic subjects showed striking inventiveness in finding some cause in the physical world to explain the speaker's literally false utterance; one subject explained the utterance in the pretend story about playing ships ("you're standing in the sea!") by saying that the boys had flooded the kitchen, and explained the white lie about being glad to receive encyclopedias instead of a rabbit as being "because the book was all about rabbits." Another subject responded to the figure of speech "a frog in your throat" by saying that the story character had swallowed a frog. A 24-year-old man with a verbal IQ of 100 explained the story about pretending a banana is a telephone by saying, "some cordless telephones are made to look like fruit."

These responses give an immediate sense of the autistic person's idiosyncratic view of events and the relative difficulty for them of attributing mental states, which makes constructing an elaborate and unusual physical explanation the preferred or easier (or perhaps the only) option. These idiosyncratic responses may also reveal peculiarities in the accessibility of different processing contexts, since they appear to normal communicators to be far more "costly" to process than the correct mental state answer. This issue is discussed, in relation to Frith's (1989) theory of a deficit in "central coherence" in autism, in Happé (in preparation b).

Striking in the results from the Strange Stories was that the autistic subjects did not use significantly fewer mental state justifications overall than the control groups. Tager-Flusberg (1989, 1992) found that young autistic children just developing language used fewer cognitive state terms in spontaneous communication than MLU-matched Down syndrome subjects, but were no different in their use of desire, perception, or emotion terms. What distinguished the autistic subjects in the present study was not a failure to use mental state terms (including cognitive terms like "think") but a failure to use the *appropriate* mental state terms in response to the Strange Stories. The fact that autistic people appear to recognize that these stories require answers in the realm of mental state language is itself surprising and intriguing. It may be that these relatively able autistic subjects have learned that there is a class of words that apply to events that are puzzling (for them), and where literal meaning of a speaker's utterance does not make sense. It would be interesting to explore the class of situations to which autistic people apply mental state terms, as suggested above for the repetitive, set explanations used by the autistic subjects in the no-ToM group.

The results from the Strange Stories, then, suggest that there are real underlying differences between the autistic subjects in the no-, 1st-order, and 2nd-order ToM groups. This supports the validity of the tasks in the ToM battery, indicating that they may measure an underlying competence in attributing mental states that is also tested by the Strange Stories. In addition the stories seem to have revealed impairments in social understanding in even the most able autistics, who passed all the tasks in the ToM battery. This may be because some of the stories were actually understood by normal subjects at a third-order theory-of-mind level (e.g., double bluff; he *knows* they *think* he will *lie*). Alternatively, it may be that the slightly more naturalistic format of the stories, and the absence of test questions drawing attention to salient elements (which are a feature of the ToM battery tasks), may reveal the difficulties even the most able autistic individuals appear to have in *applying* what social knowledge they may have in everyday life.

Happé (in preparation b) suggests one possible reason for this failure to apply a theory of mind in terms of an additional and persistent deficit in "central coherence." Autistic individuals are hypothesized to have a specific impairment in extracting meaning in context, and a preference for processing local versus global information (Frith, 1989). In this respect, the autistic subjects' large number of incorrect mental state answers is of interest. These may be taken as context-inappropriate answers, that is, answers that fit the utterance in isolation but not in the story context given. So, for example, an autistic subject who explains a white lie as a joke, may be failing to use story context to inform his answer, focusing instead on the utterance alone and the fact that it is not true (as established by specific questioning). A deficit in central coherence, then, would explain more clearly than a ToM impairment the specific pattern seen (i.e., that autistic subjects tend to give incorrect mental state answers rather than either giving no answer or resorting to purely behavioral or physical state answers). In this way, the autistic subjects' poor performance on the stories may resemble the failure of right hemisphere-damaged patients on tasks of story integration, joke completion, and interpretation of nonliteral utterances (Brownell, Michel, Powelson, & Gardner, 1983; Bryan, 1988; Kaplan, Brownell, Jacobs, & Gardner, 1990). The fact that even those autistic subjects who passed all the theory of mind tests showed characteristic inappropriate mental state answers, may suggest that a deficit in central coherence is a more universal or persistent impairment in autism than the inability to attribute mental states alone. Further exploration of these two cognitive deficits and their relation will be important for our understanding of the behavioral phenotype of autism, for remediation and education, and for the issue of subgroups within the autistic continuum.

APPENDIX

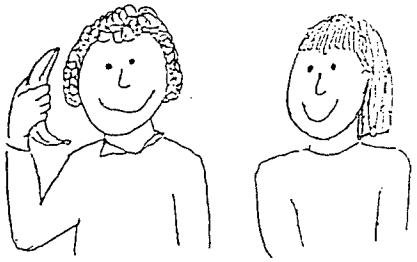
Examples of Strange Stories[3]

Story Type: Pretend

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?



Story Type: Joke

Today James is going to Claire's house for the first time. He is going over for tea, and he is looking forward to seeing Claire's dog, which she talks about all the time. James likes dogs very much. When James arrives at Claire's house Claire runs to open the door, and her dog jumps up to greet James. Claire's dog is huge, it's almost as big as James! When James sees Claire's huge dog he says, "Claire, you haven't got a dog at all. You've got an elephant!"

Is it true, what James says?

Why does James say this?

[3]The full set of test stories can be obtained from the author.

Story Type: Lie

One day, while she is playing in the house, Anna accidentally knocks over and breaks her mother's favorite crystal vase. Oh dear, when mother finds out she will be very cross! So when Anna's mother comes home and sees the broken vase and asks Anna what happened, Anna says, "The dog knocked it over, it wasn't my fault!"

Was it true, what Anna told her mother?

Why did she say this?

Story Type: White Lie

Helen waited all year for Christmas, because she knew at Christmas she could ask her parents for a rabbit. Helen wanted a rabbit more than anything in the world. At last Christmas Day arrived, and Helen ran to unwrap the big box her parents had given her. She felt sure it would contain a little rabbit in a cage. But when she opened it, with all the family standing round, she found her present was just a boring old set of encyclopedias, which Helen did not want at all! Still, when Helen's parents asked her how she liked her Christmas present, she said, "It's lovely, thank you. It's just what I wanted."

Is it true, what Helen said?

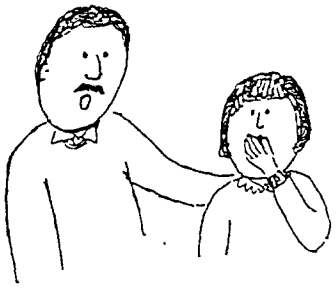
Why did she say that to her parents?

Story Type: Figure of Speech

Emma has a cough. All through lunch she coughs and coughs and coughs. Father says, "Poor Emma, you must have a frog in your throat!"

Is it true, what Father says to Emma?

Why does he say that?



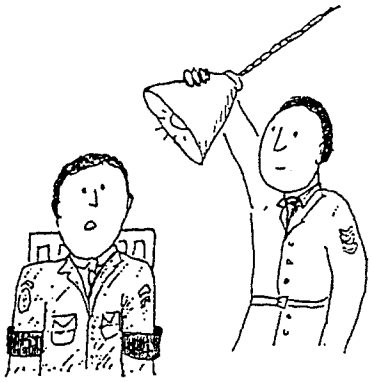
Story Type: Double Bluff

During the war, the Red army captured a member of the Blue army. They want him to tell them where his army's tanks are; they know they are either by the sea or in the mountains. They know that the prisoner will not want to tell them, he will want to save his army, and so he will certainly lie to them. The prisoner is very brave and very clever, he will not let them find his tanks. The tanks are really in the mountains. Now when the other side asks him where his tanks are, he says, "They are in the mountains."

Is it true what the prisoner said?

Where will the other army look for his tanks?

Why did the prisoner say what he said?

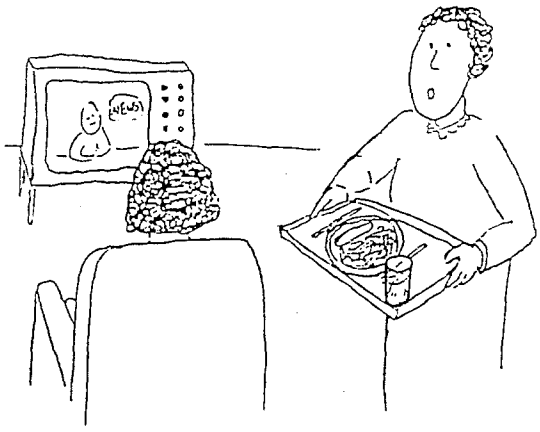


Story Type: Irony

Ann's mother has spent a long time cooking Ann's favorite meal; fish and chips. But when she brings it in to Ann, she is watching TV, and she doesn't even look up, or say thank you. Ann's mother is cross and says, "Well that's very nice, isn't it! That's what I call politeness!"

Is it true, what Ann's mother says?

Why does Ann's mother say this?



Story Type: Persuasion

Jill wanted to buy a kitten, so she went to see Mrs. Smith, who had lots of kittens she didn't want. Now Mrs. Smith loved the kittens, and she wouldn't do anything to harm them, though she couldn't keep them all herself. When Jane visited she wasn't sure she wanted one of Mrs. Smith's kittens, since they were all males and she had wanted a female. But Mrs. Smith said, "If no one buys the kittens I'll just have to drown them!"

Was it true, what Mrs. Smith said?

Why did Mrs. Smith say this to Jane?

Control Physical Story Example

Sally is in the garden. She is sowing seeds, so that next year she will have lots of vegetables in her garden. She sows seeds for carrots, lettuces and peas. She sows the seeds well, but when she goes inside after sowing them, the birds fly down and eat up all Sally's seeds! Poor Sally, not one of her seeds is left!
Q: Is it true that Sally sowed seeds for turnips and swedes?
Q: Why will Sally not have any vegetables in her garden?

Examples of Subjects' Answers to the "Why" Question

Answers Rated as Mental State Justifications

Because he doesn't like the dentist
She's cross
He's lying
Said it to fool her
She's just pretending
He's making a joke
To make them happy
It's just an expression people use
She thought it was a telephone
He knows they won't believe him
She doesn't want to upset them

Answers Rated as Physical State Justifications

So he won't have to go to the dentist
So she won't get spanked
That's where it is
Because it looks like a telephone
In order to sell the kittens
Because the dog is big
Because she won the competition

REFERENCES

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

Bowler, D. M. (1992). Theory of mind in Asperger's syndrome. *Journal of Child Psychology and Psychiatry, 33,* 877-893.

Brownell, H. H., Michel, D., Powelson, J., & Gardner, H. (1983). Surprise but not coherence; sensitivity to verbal humour in right hemisphere patients. *Brain and Language, 18,* 20-27.

Bryan, K. L. (1988). Assessment of language disorders after right hemisphere damage. *British Journal of Disorders of Communication, 23,* 111-125.

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

Frith, U., Morton, J., & Leslie, A. M. (1991). The cognitive basis of a biological disorder: Autism. *Trends in Neuroscience, 14,* 433-438.

Happé, F. G. E. (1991). The autobiographical writings of three Asperger Syndrome adults: Problems of interpretation and implications for theory. In U. Frith (Ed.), *Autism and Asperger syndrome.* Cambridge: Cambridge University Press.

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

Happé, F. G. E. (in preparation a). Does the autistic child ever have a theory of mind? Explaining the success of the "talented minority."

Happé, F. G. E. (in preparation b). The homograph task: A test of weak central coherence in autism.

Hobson, R. P. (1989). Beyond cognition: A theory of autism. In G. Dawson (Ed.), *Autism: Nature, diagnosis and treatment.* New York: Guilford.

Hobson, R. P. (1990). On acquiring knowledge about people, and the capacity to pretend: A response to Leslie. *Psychological Review, 97,* 114-121.

Kaplan, J. A., Brownell, H. H., Jacobs, J. R., & Gardner, H. (1990). The effects of right hemisphere damage on the pragmatic interpretation of conversational remarks. *Brain and Language, 38,* 315-333.

Leslie, A. M. (1987). Pretence and representation: The origins of "Theory of Mind." *Psychological Review, 94,* 412-426.

Leslie, A. M. (1988). Some implications of pretence for mechanisms underlying the child's theory of mind. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing Theories of Mind.* New York: Cambridge University Press.

Leslie, A. M., & Frith, U. (1988). Autistic children's understanding of seeing, knowing, and believing. *British Journal of Developmental Psychology, 6,* 315-324.

Ozonoff, S., Rogers, S. J., & Pennington, B. F. (1991). Asperger's syndrome: Evidence of an empirical distinction from high-functioning autism. *Journal of Child Psychology and Psychiatry, 32,* 1107-1122.

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.

Sodian, B., & Frith, U. (1992). Deception and sabotage in autistic, retarded and normal children. *Journal of Child Psychology and Psychiatry, 33,* 591-605.

Tager-Flusberg, H. (1989). *An analysis of discourse ability and internal state lexicons in a longitudinal study of autistic children.* Paper presented at the Biennial meeting of the Society for Research in Child Development, Kansas City, MO.

Tager-Flusberg, H. (1992). Autistic children's talk about psychological states: deficits in the early acquisition of a theory of mind. *Child Development, 63,* 161-172.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: representation and the constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103-128.