Social understanding: How does it fare with advancing years?

Susan Sullivan\(^1\)\* and Ted Ruffman\(^1,2\)

\(^1\)University of Sussex, UK
\(^2\)University of Otago, Dunedin, New Zealand

Until recently, theory of mind abilities have received little attention beyond the childhood years. However, pioneering work carried out by Happé, Winner, and Brownell (1998) has opened the doors on a new and exciting area of research that examines theory of mind abilities in later years. Happé et al. reported that theory of mind performance was superior in the elderly. Yet, in direct contrast to these findings, Maylor, Moulson, Muncer, and Taylor (2002) report a decline in theory of mind abilities with advancing years. We used Happé et al.'s task and, like Maylor et al., found a decline in theory of mind abilities in the elderly. Yet this deficit was related to a decline in fluid abilities. We then examined whether deficits in social understanding in the elderly could also be independent of fluid abilities. We used two new tasks: identifying emotions from still photos and identifying emotions and cognitions from video clips. Again we found a decline in social understanding in the elderly, and in this case, the decline was independent of changes in fluid abilities.

In this study we examine the social understanding of elderly individuals, using a range of tasks that tap different aspects of social cognition, including recognition of emotional and cognitive states and theory of mind abilities. Part of the driving force behind the current research was that, to date, most empirical studies have focused on theory of mind performance in infants and young children, with little attention being paid to the development of such abilities beyond the school years. In a recent attempt to redress the balance, Happé, Winner, and Brownell (1998) carried out pioneering research that examined theory of mind abilities in old age. They gave participants two types of passages: theory of mind stories and control stories. The theory of mind stories required an inference about the characters’ thoughts and feelings. In most cases, this required second-order reasoning (i.e. a consideration of what someone thinks about another person’s thoughts). For example, one of the theory of mind stories relates the tale of a

\* Correspondence should be addressed to Dr Susan Sullivan, MRC Research Studentship, Experimental Psychology, University of Sussex, Falmer, Brighton BN1 9QG, UK (e-mail: susansu@biols.susx.ac.uk).
burglar who has just robbed a shop. As the burglar is leaving the scene of his crime, a policeman sees him drop his glove. The policeman does not know that he is a burglar and just wants to tell him he has dropped his glove. But when the policeman shouts out to him the burglar gives himself up. The participant is then asked why the burglar does that. Thus, in the above example the participant must consider what the burglar is thinking about the policeman’s thoughts (i.e. second-order reasoning).

In comparison, the control stories required an inference about physical causality. For example, the participant is told about a man who is going to buy an expensive new car. However, he is considering whether to make a single payment, or whether to spread the cost over the year. If he pays in monthly instalments, the dealer will charge 5% interest on the loan. His bank currently gives him 8% interest on the money in his account. However, even though he easily has enough money to pay the full amount, he decides to pay by monthly instalments. The participant is then asked why he does that.

Happe et al. (1998) reported better theory of mind performance in older adults (mean age = 73 years) than in younger adults (mean age = 21 years), with no difference between the performance of the two age groups on stories that tapped knowledge of physical events. Their results led them to conclude that performance on theory of mind tasks remains intact and may even improve over the later adult years, thus coining the phrase ‘the getting of wisdom in old age’.

In contrast, a more recent study conducted by Maylor, Moulson, Muncer, and Taylor (2002) used stimuli based on Happe et al. (1998), and compared three age groups (mean ages = 19, 67 and 81 years) on ten theory of mind and five physical stories. Many of the passages were the same as those used in Happe et al’s (1998) study, whereas others were new. In line with Happe et al, they reported no significant age differences in performance on the physical stories. However, contrary to Happe et al., they found that theory of mind abilities were impaired in both elderly groups under task conditions comparable to those of Happe et al. In an additional condition, Maylor et al. reduced the memory demands of the stories by including cartoon pictures with five of the theory of mind stories to aid understanding of the sequence of events, and by allowing participants to continually refer back to the story text when answering the test question. They found that this procedure enabled the group aged 67 years to perform at the same level as the younger participants (though unlike Happe et al., not better). Nevertheless, the group aged 81 years continued to be impaired.

The first purpose of the present study was to help clarify the debate concerning the possible improvement or deterioration of theory of mind abilities in old age. To this end we used Happe et al’s stimuli. This is an essential starting point to determine whether we would replicate Happe et al’s findings using the same stimuli.

Second, we wanted to rule out any possibility that the impaired performance of the elderly found in Maylor et al. (2002) was due to the onset of dementia (which they did not test for). From the outset we believed this to be an unlikely cause for the impaired performance of the elderly in that study, especially as those authors excluded those with a definite diagnosis of dementia. However, in order to safeguard against this risk in the current study, all participants aged over 60 years were screened for signs of dementia using the Mini Mental State Examination (MMSE: Folstein, Folstein, & McHugh, 1975). Furthermore, all elderly participants were assessed for past history of strokes, as right-hemisphere strokes have been shown to cause pragmatic and social difficulties (Happe, Brownell, & Winner, 1999).

Third, whereas Happe et al. (1998) did not include any tests of participants’ fluid or crystallized abilities, we examined the relationship between ageing, theory of mind,
Crystallized intelligence, and fluid intelligence. Cognitive ageing is a selective process that affects different mental functions in different ways. For example, those functions associated with greater mental effort, novelty and information complexity (i.e. fluid abilities) are significantly more affected by the ageing process than crystallized abilities that rely on previously acquired expertise (Salthouse, 2000). Interestingly, an examination of the stories used in Happé et al.'s study reveals that the theory of mind stories are complex in nature and involve the manipulation of multiple propositions (Maylor et al., 2002), and correct responses on these stories requires recursive thinking (i.e. what someone thinks about another person's thoughts). Thus, it may be that Happé et al.'s theory of mind stories place a greater load on working memory than the physical stories, and that changes in working memory and general fluid abilities, rather than theory of mind understanding per se, are responsible for the decline in theory of mind task performance observed by Maylor et al.

If such arguments are correct, then one might expect that fluid ability would correlate with the theory of mind stories, particularly in the elderly (given the decline in fluid abilities in this group), but to a lesser extent with the physical stories. In addition we expected that crystallized ability would correlate with both types of stories for each age group because (a) both story types are heavily linguistically based, (b) theory of mind ability in children has been closely linked with language ability (e.g. Astington & Jenkins, 1999; Happé, 1995; Jenkins & Astington, 1996; Ruffman, Slade, Rowlandson, Rumsey, & Garnham, 2003), and (c) the test of crystallized ability taps language understanding.

In addition, unlike in previous research, we asked participants memory questions that required them to recall key parts of each passage. The memory questions would make it possible to determine whether failure on a test question was due to difficulty in recalling key parts of the text, or to a real deficit in understanding.

The fourth aim concerned how to control for the working memory requirements in test materials. Maylor et al. (2002) attempted to reduce the working memory requirements in the theory of mind stories by allowing participants to study the stories at the time they were asked the test questions. A working memory task is one that simultaneously taps both memory and processing (e.g. Case, 1985; Roberts & Pennington, 1996). Maylor et al.'s manipulation clearly reduces the working memory demands, but these demands nevertheless remain substantial (i.e. second-order reasoning). Furthermore, even if one proposition can be maintained through reading, other propositions must still be remembered and compared to the 'just read' proposition to make sense of the story. The bottleneck in working memory means that there is a limit to the amount of information that can be on-line at one time.

Therefore, the fourth aim of the study was to examine social understanding in tasks that were not heavily linguistically based and in which the working memory requirements were kept to a bare minimum. Tasks tapping different aspects of social understanding are of interest for another reason. Although stories about social events are certainly relevant to social understanding, there is an important sense in which they are lacking in ecological validity. For this reason we developed a new measure of social understanding, the video task. This task made use of moving images in an attempt to mirror the real demands of live social situations. Moving images of social interactions are likely to include a wealth of additional information relative to story-based tasks or social reasoning tests that make use of still photographs. In everyday social interactions we make use of many forms of information, such as changes in the eye region, the mouth region, and body posture, as well as general contextual information. These additional
cues might conceivably aid in the recognition of emotions. Alternatively, because the interactions are fast-moving, this might result in fleeting glimpses of facial expressions, obscuring the ability to identify the emotions portrayed. Either way, our video task would provide a measure that mirrored real-life social situations in a more naturalistic and ecologically valid way than previous tasks of this nature. The video task required the ability to understand what someone was thinking or feeling, for instance, when a person was uncomfortable, bored, confused, or lying.

In addition to the video task, we also asked participants to match emotion labels (happy, sad, afraid, angry, surprised and disgusted) with still photographs of faces. The emotion-labelling task would allow us to determine whether any social deficits found in the elderly group would be broad enough to lead to shortfalls, not just on theory of mind stories and in fast-moving social interactions (the video task), but in basic emotion recognition skills as well. Impairments in judging emotions and cognitions in social encounters, or on still faces, would be likely to have an important negative impact on one’s ability to successfully interrelate with others in everyday life.

A strength of the video and emotion-labelling tasks is that they reduce the working memory demands substantially in that there is no need to read text, combine propositions, or store the answer. The possible answer choices were available (on the video screen or on the table around the still photograph) when the participant answered the test question. Further, correct performance on the video and emotion-labelling tasks do not require second-order reasoning abilities (i.e. inferences about what one person thinks about another person’s thoughts). Instead, they require relatively simple inferences about a single person’s thoughts or feelings. Deficits on these tasks would provide evidence for a social competence impairment that was more independent of language and working memory than the tasks used in previous research and, arguably, would be more reflective of the sorts of insights that people are required to make in most social interactions. Due to the differences in fluid requirements between the various tasks, we expected that performance on each task might not correlate with one another, especially for older individuals who have degraded fluid abilities and are, therefore, more likely to be affected by fluid demands.

Finally, we examined whether there were gender differences in social understanding on our tasks. Typically, females tend to do better on social understanding, emotion recognition and false belief tasks (e.g. Bosacki, 2000; Charman, Ruffman, & Clements, 2002; Dunn, Cutting, & Demetriou, 2000; Sogon & Izard, 1987), although sometimes the advantage is very small.

**Method**

**Participants**

There were 48 participants, all of whom were reimbursed for their time. There were 24 healthy elderly individuals, ranging in age from 60 to 82 years (mean age = 73 years; 16 women and 8 men), who were compared with 24 younger participants ranging in age from 20 to 46 years (mean age = 30 years; 11 women and 13 men). All participants were from predominantly white, middle-class backgrounds, as assessed by pre-test interviews.

All of the elderly participants were recruited through a university database and all lived independently. All participants aged over 60 years were screened for signs of dementia using the Mini Mental State Examination (MMSE: Folstein et al., 1975).
used a cut-off point of 26 because Folstein et al. had demonstrated that the mean score for young adults is 27.6. No participants were excluded on this basis. The majority of the younger participants were university students or their friends. The first language of all participants was English.

Participants were given the AH4 (Heim, 1970) as a measure of fluid intelligence, and the NART (National Adult Reading Test: Nelson & Willison, 1991) as a measure of crystallized intelligence. The AH4 incorporates measures of verbal analogies, verbal opposites, arithmetical computations and synonyms. Successful performance on this task is thought to require ‘creative answer’ (i.e. fluid) solutions. The NART requires reading ‘irregularly’ spelled words. These irregular words can only be read correctly if the participant has encountered them before, and in this way the test relies on previously acquired (i.e. crystallized) knowledge. Example questions from the AH4 and NART can be found in the Appendix. Table 1 shows the mean ages and scores for the NART and AH4 intelligence measures across the two age groups. It reveals a typical cognitive profile of ageing with a greater decline in fluid abilities.

One-way analyses of variance examining fluid and crystallized abilities in the two age groups revealed significant effects of age group for the fluid intelligence measure, $F(1, 46) = 10.08, p = .003$, but not for the crystallized intelligence measure, $F(1, 46) = <1.00, \text{n.s.}$ Thus, in line with the cognitive ageing literature, the measures of fluid and crystallized intelligence revealed the typical pattern of ageing, with the younger participants superior only on the fluid measure, a finding in keeping with the notion that fluid but not crystallized intelligence declines with advancing years (Salthouse, 2000).

### Materials

The theory of mind and physical stories consisted of 12 short passages of text (6 theory of mind and 6 physical stories). They were adapted from Happé’s (1994) *Strange stories*, and were identical to those used in Happé et al.'s (1998) study, although in that study 16 passages (8 theory of mind and 8 physical stories) were utilized. They were also the same passages used by Maylor et al. (2002), although these authors also created new passages of their own. Each story was followed by a test question about a character’s thoughts, feelings and intentions or about physical causality. Each of the passages was printed on a piece of paper. After participants had read the text they turned over the page to reveal the test question. Following their response they turned over to the next page where the three memory questions were printed. The order of presentation of the story types (i.e. theory of mind vs. physical) was counterbalanced across

<table>
<thead>
<tr>
<th></th>
<th>Elderly</th>
<th>Young</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>73 (6.0)</td>
<td>30 (7.5)</td>
</tr>
<tr>
<td>AH4 (percentage correct)</td>
<td>87.4 (8.3)</td>
<td>93.4 (4.0)</td>
</tr>
<tr>
<td>NART (standard score)</td>
<td>117.2 (7.3)</td>
<td>118.4 (4.9)</td>
</tr>
</tbody>
</table>
participants. Examples of theory of mind and control stories are included in the Appendix.

The video task initially consisted of 41 video clips. These clips were given to 14 adults aged between 21 and 65 years in a pilot study (mean age = 52 years). On the basis of this data, each clip was ranked according to level of difficulty. Out of the 41 clips, 26 were chosen for the final video task. These 26 were chosen on the basis that the target word was selected in preference to the foil at a greater than chance rate. Thus, we required 11 of 14 participants to select the target word (binomial Test: \( k = 11, N = 14, p < .05 \)). These 26 clips were then split into two groups of 13, which were recorded onto separate video cassettes (i.e. video 1 and video 2). The clips on the two videos were matched for level of difficulty. Each participant was tested using either video 1 or video 2 such that an equal number of participants in each age group watched each video. The Appendix provides a complete description of the target and foil words used in the video task.

In many studies, foil words are the semantic opposite of the target word, for instance, a serious vs. a playful message (Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997). However, in the video task, we thought that using semantic opposites would make the task too easy. Thus, the foil words were initially chosen by three judges on the basis that they could be plausible, but less accurate labels for the mental state exhibited in the scene, especially for those participants who might demonstrate theory of mind deficits on Happé’s (1994) Strange stories.

The main character in each video clip was sometimes male and sometimes female. Approximately half of the clips featured actors ranging in age from 18 to 26 years, whereas the remaining clips featured more mature actors ranging in age from 40 to 70 years. Clips were taken from various television programs, news clips and movies. For example, in one video clip the participant saw a young woman (approximately 30 years old) sitting on a sofa flicking through a magazine. On the basis of her body posture and facial expression the participant’s task was to decide whether she was ‘bored’ or ‘sad’. Each clip was between 2 and 7 seconds long. For 3 seconds before the clip appeared on the screen, two mental state terms were presented on the screen, one at the bottom left-hand side and the other at the bottom right-hand side. These mental state terms remained on the screen as the clip played. The locations of the target word and the foil were counterbalanced as to whether they appeared on the bottom left side or the bottom right side of the monitor. Participants were informed beforehand that their task was to determine which of the two words that appeared on the screen best described what the character in each clip was thinking or feeling.

The emotion-labelling task was based on Ekman (1992) and required participants to match emotion labels with still photographs of faces. Six emotion labels were used: happy, sad, afraid, angry, surprised and disgusted. The labels were always available to the participant on the table around the still photographs so as to reduce the fluid requirements for the task. The participants were instructed to match the correct label with the correct photograph in any order.

**Procedure**

All participants were tested individually in the same quiet university laboratory at the University of Sussex, UK. For the ‘strange stories’, they were presented with a test folder that included all six theory of mind stories or all six physical stories. There were written instructions on the first page of each file, identical to those used in Happé et al.’s (1998)
study. These instructions informed the participant that they should read each of the following passages silently until they had understood them, turn the page for the test question, and verbally provide their answer to the examiner. As in Happé et al.'s original study, the written instructions also informed the participants that once they had turned the page to reveal the test question they were not permitted to refer back to the text. Therefore, they were told to take as long as they deemed necessary to examine the text before turning the page. After answering the test question, they were instructed to turn to the next page to reveal the memory questions.

The participant’s answers to the test and memory questions were later scored according to the marking scheme created by Happé et al. (1998). In line with that study, the answers were awarded marks according to whether the participant was completely incorrect or did not know the answer (0), gave a partially correct answer (1), or a full and explicit correct answer (2). A second marker who was blind to both the participant group and the hypotheses coded 30% of the transcripts. The criteria for second coding were stringent in that any difference between scores was counted as a disagreement. The first and second markers had a 91% agreement rate. Disagreements were resolved by discussion. The memory questions were scored as correct or incorrect. Examples of the memory questions and scoring criteria can be found in the Appendix.

For the video task, participants were told that they were to watch a video that was made up of 13 clips, some of which were only a few seconds long. Participants were instructed that they should decide which of the two words that appeared on the screen best described the way the person in the clip was feeling. Participants could inform the experimenter of their decision either during the clip or when the clip ended. There was no indication that the elderly tended to impulsively answer the test question at an earlier time than did the young adults.

Order of presentation of the stories, emotion-labelling task, video task, and NART and AH4, were counterbalanced across all participants.

**Results**

**Strange stories**

We examined participants’ scores on the theory of mind and physical stories. Table 2 includes this information. We conducted a 2 (age group: elderly vs. young) × 2 (story type: theory of mind vs. physical) × 2 (gender) analysis of variance with story score as the dependent variable. This analysis revealed a significant main effect of story type, \(F(1, 44) = 11.27, p = .001\), and a significant interaction between age and story type, \(F(1, 44) = 4.29, p < .05\). There were no other significant effects. The pattern of means within the interaction was explored with \(t\)-tests, which revealed one significant effect. The younger group performed significantly better than the older group on the theory of mind stories, \(t(46) = 4.56, p < .001\). This finding appears to be in complete contrast to that reported by Happé et al. (1998). There was a trend for the younger group to do better on the control stories as well, although the effect failed to reach significance, \(t(46) = 1.82, p < .08\).

Next we examined reasons for why the elderly group had a particular impairment on the theory of mind stories (as indicated by the significant interaction). One possible reason for older adults’ difficulty is forgetting story information. Table 2 includes performance on the memory questions. We analysed this data with a 2 (age group: elderly vs. young) × 2 (story type: theory of mind vs. physical) analysis of variance with
memory score as the dependent variable. There was a significant effect for age group, $F(1, 46) = 17.17, p < .001$, but no effect for story type, $F(1, 46) = < 1.00$, n.s., and no interaction, $F(1, 46) = < 1.00$, n.s. Thus, although the elderly participants had poorer memory for story information, they had equal difficulty on ... were significant for the young adults. The correlation between fluid ability and (a) theory of mind performance was $r(22) = .40, p < .05$, and (b) physical story performance was $r(22) = .38, p < .05$. The correlation between crystallized ability and (a) theory of mind performance was $r(22) = .44, p < .05$, and (b) physical story performance was $r(22) = .30$, n.s.

Given that theory of mind performance correlated with fluid ability in the elderly group, we then examined whether differences in fluid intelligence could account for the differences in performance between the two age groups. We conducted a 2 (age group) × 2 (story type) analysis of covariance with story score as the dependent variable and AH4 performance as a covariate. The adjusted means for the two age groups are shown in Table 2. Once AH4 performance was accounted for, there was no longer a significant interaction between age and story type, $F(1, 45) = 2.49, p = .12$. Thus, the difference between age groups on Happé et al.’s (1998) theory of mind stories was related to fluid abilities. Further, as crystallized intelligence (NART scores) had also been

### Table 2. Mean scores and percentages for the theory of mind (ToM) and physical stories

<table>
<thead>
<tr>
<th></th>
<th>Elderly</th>
<th>Adjusted means*</th>
<th>Adjusted means**</th>
<th>Young</th>
<th>Adjusted means*</th>
<th>Adjusted means**</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToM stories:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean scores</td>
<td>7.0 (2.2)</td>
<td>7.3 (.38)</td>
<td>7.1 (.36)</td>
<td>9.5 (1.3)</td>
<td>9.2 (.38)</td>
<td>9.4 (.36)</td>
</tr>
<tr>
<td>Percentage</td>
<td>58</td>
<td>61</td>
<td>59</td>
<td>79</td>
<td>77</td>
<td>78</td>
</tr>
<tr>
<td>Physical stories:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean score</td>
<td>6.6 (2.3)</td>
<td>6.8 (.48)</td>
<td>6.6 (.45)</td>
<td>7.8 (2.1)</td>
<td>7.6 (.48)</td>
<td>7.7 (.45)</td>
</tr>
<tr>
<td>Percentage</td>
<td>55</td>
<td>57</td>
<td>55</td>
<td>65</td>
<td>63</td>
<td>64</td>
</tr>
<tr>
<td>Memory questions:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ToM</td>
<td>16.6 (1.2)</td>
<td>17.7 (0.6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage</td>
<td>92</td>
<td></td>
<td></td>
<td>98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Physical</td>
<td>16.5 (1.2)</td>
<td>17.5 (0.7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percentage</td>
<td>92</td>
<td></td>
<td></td>
<td>97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Adjusted means with fluid intelligence (AH4) as the covariate and standard errors shown in parentheses.

* Adjusted means with crystallized intelligence (NART) as the covariate and standard errors shown in parentheses.

* Maximum score = 12; d maximum score = 18. Standard deviations are shown in parentheses.
found to correlate with the theory of mind stories, we examined whether differences in crystallized intelligence could account for the differences in performance between the two age groups. We conducted a 2 (age group) \times 2 (story type) analysis of covariance with story score as the dependent variable and NART scores as the covariate. With NART scores accounted for, there was still a significant interaction between age and story type, \( F(1, 45) = 4.10, \ p < .05 \). Thus, the difference between the age groups on Happé et al.'s theory of mind stories was more a factor of fluid than crystallized ability.

Then, in order to allow for a more direct comparison with Maylor et al.'s (2002) study, the older participants were split into two age groups. These were young-old (mean age = 68 years, \( N = 12, \) range = 60–73 years), and old-old (mean age = 77 years, \( N = 12, \) range = 74–82 years). We examined this data with a 2 (age group: young-old vs. old-old) \times 2 (story type: theory of mind vs. physical) analysis of covariance, with story score as the dependent variable and AH4 performance as a covariate. There was no main effect of age, \( F(1, 21) < 1.00, \) n.s. Nor was there a main effect of story type, \( F(1, 21) < 1.00, \) n.s., or an interaction between age and story type, \( F(1, 21) < 1.00, \) n.s.

Thus, there were no differences in performance between the young-old group and the old-old group. This pattern of results is consistent with the idea that theory of mind performance drops off some time in the seventh decade and remains at that level into the eighth or ninth decade.

Finally, we examined whether there was a subset of elderly participants who exhibited unimpaired social understanding. To this end, we computed the proportion of elderly individuals who scored above the younger group mean for each task. For the theory of mind stories, only 13% of elderly individuals appeared to have unimpaired social understanding as tapped by these stories, with this figure rising slightly for the physical stories to 25% of the elderly scoring above the mean for the younger group.

**Video and emotion tasks**

Planned comparisons on the video task revealed that the participants who received the first set of video clips (video 1) scored equally to those who received the second set of clips (video 2). The mean on video 1 was 10.71 (SD = 1.37), whereas the mean on video 2 was 10.25 (SD = 1.54), \( t(46) = 1.09, \) n.s. This confirms that the two videos were successfully matched for level of difficulty, and for this reason we collapse the results for this variable in subsequent analyses.

The mean scores for the video and emotion tasks are included in Table 3. Similar to the 'strange stories' task, we examined whether fluid or crystallized intelligence could account for the differences in performance between the two age groups. Therefore, scores for the video and emotion-labelling tasks were converted to percentage correct and the data were analysed with a 2 (task: video vs. emotion-labelling) \times 2 (age group) \times 2 (gender) analysis of covariance, with scores on the AH4 (fluid intelligence) and the NART (crystallized intelligence) as covariates, and percentage correct on the video task and emotion-labelling task as the dependent variable. The adjusted means are shown in Table 3. Unlike on the strange stories, the younger group was significantly better on the video and emotion task after accounting for both AH4 and NART performance. Thus, we found a main effect of age, \( F(1, 42) = 8.65, \ p < .01, \) but no interaction between age and emotion task, \( F(1, 42) = 1.23, \) n.s. The only other significant effects were an interaction between age and gender, \( F(1, 42) = 7.28, \ p < .05, \) and between task and gender, \( F(1, 42) = 6.11, \ p < .05. \) The former interaction was uninteresting in that it simply indicated that the proportion of males to females was reversed in the two age groups.
The interaction between task and gender was explored with t-tests that revealed one marginally significant effect. Males were marginally worse on the emotion-labelling task than females, \( t(46) = -1.98, p < .06 \) (male mean = 81.8%, SD = 15.6%; female mean = 89.4%, SD = 11.5%).

<table>
<thead>
<tr>
<th>Elderly:</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean score</td>
<td>9.9 (1.4)</td>
<td>9.8 (0.2)</td>
<td>4.9 (0.93)</td>
</tr>
<tr>
<td>Mean percentage correct</td>
<td>76 (10.3)</td>
<td>75.5 (2.3)</td>
<td>82 (15.4)</td>
</tr>
<tr>
<td>Young:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean score</td>
<td>11.0 (1.4)</td>
<td>10.5 (0.38)</td>
<td>5.4 (0.65)</td>
</tr>
<tr>
<td>Mean percentage correct</td>
<td>85 (10.4)</td>
<td>81 (2.9)</td>
<td>90 (10.9)</td>
</tr>
</tbody>
</table>

Note. Standard deviations are shown in parentheses.

Adjusted means with fluid intelligence (AH4) and crystallized intelligence (NART) as covariates, and standard errors shown in parentheses.

There was converging evidence for the lack of a mediating effect for fluid and crystallized ability in that AH4 and NART performance did not correlate with scores on the emotion-labelling or video task in either age group. In the young group, all correlations were less than, \( r = .21 \), and in the elderly group they were all less than, \( r = .07 \).

As for the strange stories, we then checked whether the decline in the elderly group was primarily due to the older half. Thus, the elderly participants were split into young-old and old-old as before. When their percentage scores were collapsed across both the video task and the emotion recognition task, the old-old group (79% correct, SD = 10%) did only slightly and non-significantly worse (as measured by t-tests) than the young-old group (80% correct, SD = 8%). Thus, the lower performance of the elderly came not just from the older half, but from the entire group.

As before, we then examined whether there was a subset of elderly participants who exhibited unimpaired social understanding. Again, we computed the proportion of elderly individuals who scored above the younger group mean for each task. For the emotion-labelling and video tasks, 33% and 29% of the elderly group, respectively, demonstrated unimpaired abilities.

**Inter-task correlations**

We computed inter-task correlations separately for each age group. For the younger group, there was a correlation between performance on the emotion-labelling task and the theory of mind stories: \( r = -.43, p < .05 \). For the older group, there was a correlation between performance on the theory of mind and physical stories: \( r = .65, p < .01 \).

**General discussion**

One of the aims of the current study was to determine whether theory of mind abilities improved or deteriorated with advancing years. At first glance the present findings
appear to be in keeping with those of Maylor et al. (2002), who found a decline in theory of mind skills with old age. However, once fluid intelligence was accounted for, there was no longer a significant difference between the two age groups. This result is consistent with the idea that an age-related decline in social understanding, as tapped by Happé et al.'s (1998) Strange stories, may be partially mediated by declining fluid intelligence. It is also consistent with findings that 3- and 4-year-olds' verbal performance on a false belief task correlates with a measure of fluid ability, working memory (Davis & Pratt, 1995; Gordon & Olson, 1998; Keenan, Olson, & Marini, 1998). It is plausible that a deficit in social reasoning, even if it is sometimes mediated by fluid abilities, will nevertheless have implications in real-world social settings, so that situations that place high demands on fluid abilities will result in impaired social understanding. In this way, Happé et al.'s task provides a useful indicator of how elderly individuals' theory of mind skills would fare under the heavy demands of fast-moving everyday social interactions.

In line with both Maylor et al. (2002) and Happé et al. (1998), there was no difference between the groups on the physical stories, although the effect in our study approached significance. However, the presence of the interaction, and the finding that the difference between the age groups on the control stories is approximately half of what it is on the theory of mind stories (1.2 vs. 2.5), is consistent with the idea that theory of mind reasoning is more adversely affected by the ageing process relative to reasoning about non-social issues. In sum, our results contradict those of Happé et al., who found that theory of mind performance improved with age (mean scores: young = 12.8, old = 14.9, maximum score possible = 16), but are more in keeping with those of Maylor et al., who also found a decline in theory of mind abilities in the elderly (mean scores: young = 4.04, old-old = 2.48, maximum score possible = 7).

As stated at the outset, neither we nor Maylor et al. (2002) have managed to eliminate the heavier working memory load in the theory of mind stories. The theory of mind stories still require second-order reasoning abilities, whereas the physical stories do not. This was one reason for including two measures of social reasoning that placed reduced demands on working memory. Using these new tasks, we found a decline in social understanding in the elderly even after co-varying out fluid and crystallized abilities. In addition, these tasks (particularly the video task) tap insights that, arguably, are more characteristic of the ‘real world’ than story-based tasks. Thus, even in situations with low working memory demands, elderly people tend to have impaired social understanding.

With regard to sex differences, we found a female advantage on one of the three tasks we used, although this effect was only marginally significant. This finding is largely in keeping with the studies reviewed earlier, demonstrating a slight female advantage on such tasks.

Our predictions regarding correlations with fluid and crystallized abilities were only partially supported. As expected, the elderly group’s performance on the theory of mind stories was related to fluid ability and crystallized ability. Unexpectedly, however, their performance on the physical stories also correlated with fluid ability. This is most likely to have been due to the way in which the physical stories were written. As for the theory of mind stories, there was a need to integrate story material in the physical stories and this would have created some demands on fluid ability, even if the theory of mind stories made an additional demand on recursive thinking (thinking about thoughts about thoughts). Also unexpectedly, crystallized ability was not related to the young adults’ performance on either story type. This may simply have been because of the restricted variance on the fluid and crystallized tests in the young adults relative to the elderly participants. Consistent with expectations, performance on the video and
emotion-labelling tasks was not related to fluid ability. Nor was it related to crystallized ability.

The age-related deficits in social understanding found in the present study are perhaps surprising in view of findings of adaptive social cognition in the elderly population (e.g. Blanchard-Fields & Chen, 1996). However, our findings fit well with those of other studies demonstrating age differences in the correction of social judgments (Chen & Blanchard-Fields, 2000), a general decline in socio-cognitive reasoning with advancing years (Pratt, Pratt, Diessner, Hunsberger, & Pancer, 1996), and a decline in the ability to identify affect with age (McDowell, Harrison, & Demaree, 1994).

There is also some research indicating that elderly individuals exhibit a deficit in the labelling of negative emotions, especially ‘sad’ (Chen & Blanchard-Fields, 2000; McDowell et al., 1994). In the current study, the elderly participants did not do worse on sadness (see Tables 4 and 5). Furthermore, any breakdown in terms of ‘negative’ emotions may be counterproductive, as such umbrella terms obscure finer details that may ultimately be more enlightening (see below).

<table>
<thead>
<tr>
<th>Table 4. Performance on the video task by specific emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotion</td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>Relieved</td>
</tr>
<tr>
<td>Excited</td>
</tr>
<tr>
<td>Flirtatious</td>
</tr>
<tr>
<td>Sympathetic</td>
</tr>
<tr>
<td>Thoughtful</td>
</tr>
<tr>
<td>Concerned</td>
</tr>
<tr>
<td>Sad</td>
</tr>
<tr>
<td>Frustrated</td>
</tr>
<tr>
<td>Distrustful</td>
</tr>
<tr>
<td>Anxious</td>
</tr>
<tr>
<td>Scared</td>
</tr>
<tr>
<td>Sarcastic</td>
</tr>
<tr>
<td>Lying</td>
</tr>
<tr>
<td>Angry</td>
</tr>
<tr>
<td>Worried</td>
</tr>
<tr>
<td>Uncomfortable</td>
</tr>
<tr>
<td>Bored</td>
</tr>
<tr>
<td>Confused</td>
</tr>
<tr>
<td>Dominant</td>
</tr>
</tbody>
</table>

Note. Maximum score possible = 12 per clip.

It is not clear why Happé et al. (1998) did not find a decline in abilities whereas we and Maylor et al. (2002) did. One possibility is differences in materials. Happé et al. used eight theory of mind stories and eight physical stories. Maylor et al. used some of these stories plus some of their own (ten theory of mind stories and five physical stories). We used six of Happé et al.’s theory of mind stories and six of their physical stories. There is
no reason for thinking that the subset of stories that we used or that the stories used by Maylor et al. were inferior to those used by Happé et al. Hence, we do not think this is a plausible reason for the difference in performance.

Another possibility for the difference in findings is that some of the elderly participants in the studies might have suffered from right-hemisphere strokes, which have been shown to cause pragmatic and social difficulties (Happé et al., 1999). Although Maylor et al. did not report on the incidence of strokes, the elderly participants in the current study were screened for past history of strokes and so this again seems an unlikely reason for the differences in findings. Furthermore, the majority of Maylor et al.’s elderly participants were members of a local whist club, a game that requires fast and efficient recall of cards played. It is conceivable that this game might even hone one’s theory of mind skills (e.g. understanding of bluffing) due to the practice in such skills that this particular game offers.

The finding that social understanding deficits are still found in a population that may have a relative advantage in terms of memory capabilities and the ability to predict what another person would do next, only lends further support to the findings of Maylor et al.'s study and the current research. In our view the most plausible reason for the difference in results is the samples that were tested. It is possible that Happé et al.'s (1998) sample were simply slower to show cognitive deterioration and in this sense they were a relatively gifted group of elderly people. Further evidence for this notion is that on each of our tasks there was a proportion of elderly individuals (between 13% and 33%) who exhibited unimpaired social understanding. This is consistent with the idea that at least a minority of elderly individuals will demonstrate preserved or superior theory of mind abilities.

What is clear is that there is no reason to question the validity of our results or those of Maylor et al. (2002). Indeed, whereas Happé et al. (1998) tested only 19 elderly participants, Maylor et al. tested 80, and we tested 24. The bulk of the evidence is consistent with the idea that elderly individuals will generally succumb to a decline in social understanding. Indeed, we found consistent deficits in the elderly group on three tasks tapping different aspects of social reasoning. Further, our results are validated by the fact that our participants displayed the typical cognitive profile of elderly individuals. Their fluid abilities were relatively impaired whereas their crystallized

Table 5. Performance on the emotion-labelling task by specific emotion

<table>
<thead>
<tr>
<th>Emotion/classification</th>
<th>Elderly score</th>
<th>Young score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>24</td>
<td>23</td>
</tr>
<tr>
<td>Surprised</td>
<td>23</td>
<td>24</td>
</tr>
<tr>
<td>Sad</td>
<td>21</td>
<td>22</td>
</tr>
<tr>
<td>Angry</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Afraid</td>
<td>18</td>
<td>17</td>
</tr>
<tr>
<td>Disgusted</td>
<td>18</td>
<td>24</td>
</tr>
</tbody>
</table>

*Note. Maximum score possible = 24.*
abilities were intact. There was nothing unusual about this group of elderly adults. Unfortunately, such information was not collected in the Happé et al. study.

Finally, we think that the pattern of results in this experiment has implications for the nature of cognitive decline in ageing. The video and emotion recognition tasks did not correlate with fluid ability. The impairment of the elderly on the video and emotion recognition tasks is consistent with the idea that changes in social understanding are not a simple function of domain-general changes in cognitive abilities. Instead, the deterioration in performance is consistent with a more domain-specific decline in social understanding that is independent from changes in fluid IQ at the very least. In this sense the results are consistent with the suggestion that there are regions of the brain dedicated to social understanding that are at least partially independent of domain-general processing (e.g. Baron-Cohen, 2000; Leslie, 1992).

For the younger group, there was a negative correlation between performance on the emotion-labelling task and the theory of mind stories. There is clearly no obvious explanation for this result other than chance. Indeed, we conducted 14 inter-task correlations in total (7 for each age group), making it plausible that chance accounted for the negative correlation. For the older group, there was a correlation between performance on the theory of mind and physical stories, possibly due to larger fluid requirements. Differences in the brain regions activated by our different social understanding tasks may also account for the lack of positive inter-task correlations.

For instance, Fletcher et al. (1995) found that performance on the ‘strange stories’ (in comparison to the control stories) activated the left medial prefrontal cortex. In contrast, emotion recognition tends to activate different areas. Sadness activates the amygdala (Blair, Morris, Frith, Perrett, & Dolan, 1999; Fine & Blair, 2000; although see Calder et al., 1996), fear activates the amygdala (Calder et al., 1996), disgust the basal ganglia and insula (Calder, Lawrence, & Young, 2001), and anger the orbito-frontal cortex (Blair & Cipolotti, 2000; Blair et al., 1999). These different regions of activation give rise to the notion that tasks tapping different aspects of social understanding draw on somewhat different insights. In this vein, fluid and working memory abilities are thought to activate the dorsolateral prefrontal cortex (Casey et al., 1995; for a summary, see Levy & Goldman-Rakic, 2000). In sum, these differences in the mediating brain region could account for the dearth of correlations between the different social understanding tasks, and between social understanding and fluid deterioration.

Thus, although our results contradict those of Happé et al. (1998) in that we found a decline in social understanding in the elderly, they are consistent in another way. Like Happé et al., our findings are consistent with the idea that social understanding is not reducible to a fluid deficit. That is, we found evidence for social understanding being modular, at least in so far as it is partially independent of fluid ability. Happé et al. were cautious in reaching this conclusion, pointing out that with no IQ measure, it was possible that their elderly group was more intelligent than their young adults. We can say with some confidence that the social understanding deficits displayed by our elderly group were not a function of fluid or crystallized intelligence. This is an essential first step in establishing modularity, though it is by no means exhaustive. For instance, stronger claims would be possible with the inclusion of an appropriate ‘non-social’ control for the video task. This is an important avenue for future research.

In sum, the results of the present study are consistent with the idea that there is a decline in social understanding in the elderly that is at least partially independent of a general decline in fluid intelligence. It is possible that there is a subset of gifted elderly people whose social understanding develops throughout the lifespan or whose social
understanding declines relatively late, so that Happé et al. (1998) did not find a decline. However, there is growing evidence—the present study plus that of Maylor et al. (2002)—that many elderly people will experience a decline in their social reasoning abilities with advancing years.

**Acknowledgements**

Many thanks to the Economic and Social Research Council for financial support to Ted Ruffman (award no. R000237071), and to the MRC (award no. G78/6293) for support to Susan Sullivan. Also many thanks to Francesca Happé for the copies of her task, and for her and an anonymous reviewer's helpful comments.

**References**


Received 8 March 2002; revised version received 19 December 2002
Appendix

**Example theory of mind story, the scoring schema and accompanying memory questions**
Simon is a big liar. Simon’s brother Jim knows this, he knows that Simon never tells the truth! Now yesterday Simon stole Jim’s ping-pong paddle, and Jim knows Simon has hidden it somewhere, though he can’t find it. He’s very cross. So he finds Simon and he says, “Where is my ping-pong paddle? You must have hidden it either in the cupboard or under your bed, because I’ve looked everywhere else. Where is it, in the cupboard or under your bed?” Simon tells him the paddle is under his bed.

**Test Question:**
Why will Jim look in the cupboard for the paddle?
Example responses scored:
2. “Jim will look in the cupboard because Simon said it was under the bed and because he’s a liar that’s not going to be true.”
1. “Because he knows Simon’s a liar.”

**Memory questions:**
(a) What does Jim know about Simon?
(b) What did Simon do yesterday?
(c) Where does Simon say the paddle is?

**Example physical story, the scoring schema and accompanying memory questions**
A burglar is about to break into a jewellers’ shop. He skilfully picks the lock on the shop door. Carefully he crawls under the electronic detector beam. If he breaks this beam it will set off the alarm. Quietly he opens the door of the storeroom and sees the gems glittering. As he reaches out, however, he steps on something soft. He hears a screech and something small and furry runs out past him, towards the shop door. Immediately the alarm sounds.

**Test question:**
Why did the alarm go off?
Example responses scored:
2. “The alarm was set off by something small and furry running out of the shop after the burglar had trodden on it.”
1. “Because something broke the beam.”

**Memory questions:**
(a) What does the burglar crawl under?
(b) What will happen if he breaks the beam?
(c) What happens when he steps on something soft?
Example questions from the AH4
1. Fish is to swim as bird is to . . . man, fly, walk, airplane, sparrow.
2. 2, 4, 8, 16, 32 . . . What number comes next?
3. Here are three figures: 5 9 4. Subtract the smallest figure from the biggest and multiply the result by the figure printed immediately before the biggest figure.
4. Young means the same as . . . youthful, ancient, vigorous, hot, baby.

Example of words included in the NART
CHORD
ACHE
DEPOT
PSALM
IDYLL
TOPIARY
BEATIFY
PRELATE

Target mental state terms and their foil terms in the video task

<table>
<thead>
<tr>
<th>Clip number</th>
<th>Target term</th>
<th>Foil</th>
<th>Clip number</th>
<th>Target term</th>
<th>Foil</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Sad</td>
<td>Angry</td>
<td>14</td>
<td>Anxious</td>
<td>Angry</td>
</tr>
<tr>
<td>2</td>
<td>Frustrated</td>
<td>Angry</td>
<td>15</td>
<td>Scared</td>
<td>Confused</td>
</tr>
<tr>
<td>3</td>
<td>Worried</td>
<td>Sad</td>
<td>16</td>
<td>Sympathetic</td>
<td>Bored</td>
</tr>
<tr>
<td>4</td>
<td>Uncomfortable</td>
<td>Worried</td>
<td>17</td>
<td>Thoughtful</td>
<td>Distrustful</td>
</tr>
<tr>
<td>5</td>
<td>Sad</td>
<td>Scared</td>
<td>18</td>
<td>Concerned</td>
<td>Interested</td>
</tr>
<tr>
<td>6</td>
<td>Bored</td>
<td>Sad</td>
<td>19</td>
<td>Sarcastic</td>
<td>Flirtatious</td>
</tr>
<tr>
<td>7</td>
<td>Lying</td>
<td>Uncomfortable</td>
<td>20</td>
<td>Worried</td>
<td>Confused</td>
</tr>
<tr>
<td>8</td>
<td>Angry</td>
<td>Frustrated</td>
<td>21</td>
<td>Angry</td>
<td>Frustrated</td>
</tr>
<tr>
<td>9</td>
<td>Confused</td>
<td>Angry</td>
<td>22</td>
<td>Anxious</td>
<td>Angry</td>
</tr>
<tr>
<td>10</td>
<td>Dominant</td>
<td>Angry</td>
<td>23</td>
<td>Relieved</td>
<td>Frustrated</td>
</tr>
<tr>
<td>11</td>
<td>Dominant</td>
<td>Angry</td>
<td>24</td>
<td>Distrustful</td>
<td>Bored</td>
</tr>
<tr>
<td>12</td>
<td>Worried</td>
<td>Frustrated</td>
<td>25</td>
<td>Sympathetic</td>
<td>Worried</td>
</tr>
<tr>
<td>13</td>
<td>Flirtatious</td>
<td>Sad</td>
<td>26</td>
<td>Excited</td>
<td>Happy</td>
</tr>
</tbody>
</table>