The automated windows task: the performance of preschool children, children with autism, and children with moderate learning difficulties

James Russell a,*, Suzanne Hala b, Elisabeth Hill c

a Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3JD, UK
b Department of Psychology, University of Calgary, Calgary, Canada
c Institute of Cognitive Neuroscience, London, UK

Received 1 February 2002; received in revised form 11 December 2002; accepted 11 December 2002

Abstract

The windows task (a hybrid executive/theory of mind task) requires participants to infer the utility of indicating a box in which a desired object is absent, adjacent to one where it is visible, in the presence of an experimenter in a completive role, and to maintain this response over a number of subsequent trials. We presented two pairs of groups (3 years vs. 4 years and autism vs. MLD) with an automated version of the task in order to find out whether the main locus of difficulty was executive or social. In two conditions the required response was to press a button beneath the chosen box, and in two conditions the deceptive or the merely competitive element was maintained. In the deceptive condition the participant had to tell the opponent which button to press. Automation removed group differences in performance on the first ‘inferential’ experimental trial, although group differences did remain in performance across all trials. We argue that it was the novel response mode used in the automated task that made it easier than the standard versions: button pressing (or indicating in the deceptive condition) rather than reference to a location. We discuss this possibility with regard to the kind of inhibitory requirements made by theory of mind tasks.

© 2002 Elsevier Science Inc. All rights reserved.

Keywords: Autism; Windows task; Executive function; Theory of mind; Pre-school

* Corresponding author. Tel.: +44-1223-333553; fax: +44-1223-333564.
E-mail address: jr111@cus.cam.ac.uk (J. Russell).

0885-2014/02/S – see front matter © 2002 Elsevier Science Inc. All rights reserved.
doi:10.1016/S0885-2014(02)00163-6
1. Introduction

Because executive tasks make a variety of inter-related demands (e.g., inhibition, working memory, self-monitoring, rule-following, instigation) it is difficult to know what significance to place upon the fact that young children are at least as challenged by them as by mentalising tasks (Perner & Lang, 1999) and that children with autism sometimes show specific deficits in the executive area (Pennington & Ozonoif, 1996; Russell, 1997). It is, in fact, important to determine the main locus of executive difficulty for these populations because tasks with a mentalistic content, such as the false belief task, sometimes make significant executive demands (Russell, Saltmarsh, & Hill, 1999; Russell, Hill, & Franco, 2001). For this reason, the development of mental concepts must be assessed in relation to executive development.

The present paper tackles this issue in the following way. We take a single task that is generally failed by 3-year-olds and passed by 4-year-olds, and on which children with autism show a specific deficit, and modify it in order to determine whether the locus of difficulty is executive or social/mentalising. This is the windows task (originally reported by Russell, Mauthner, Sharpe, & Tidswell, 1991). In its original form, the task could be construed either as an executive or as a 'mentalising' task. This is because, while requiring participants to inhibit a prepotent response and hold an arbitrary rule in working memory, it could also be said to require some understanding of deception or competition. The main aim of the current study was to eradicate entirely the deceptive and competitive elements of the task by mechanising it. If the challenge remained, this would be good evidence of its having an essentially executive nature. We also included, however, two further conditions in which some degree of deception and competition was required despite the procedure’s being mechanical in nature.

The windows task was originally designed as a non-verbal test of whether an individual can infer the utility of a deceptive act given that he or she has never experienced its utility in the past. The design of the task ensures that one can witness children producing deceptive actions for which they have not previously been reinforced and which therefore cannot be explained in terms of prior conditioning (contrast Woodruff & Premack, 1979). In the first phase of the task the child is faced with two opaque boxes and an opponent sitting opposite. On each trial one of the boxes contains a treat the location of which both child and opponent are ignorant. The child points to one of the boxes in order to tell the opponent which one to open. If this turns out to be the baited box, the opponent wins the treat. If, however, the opponent happens to have been directed to the empty box the treat goes to the child. After a number of trials, the child thereby learns that it is in his or her interest to tell the opponent that the treat is in the empty box while not knowing, prior to responding, which of the boxes is in fact empty. In the test phase the boxes are replaced by ones containing windows facing the child but not visible to the opponent. Thus, on the first test trial the child can see which of the boxes is baited and which is empty.
One can ask, first of all, whether the child can infer the strategy of pointing to the empty box on the first trial. Three-year-old children typically do not and 4-year-old children typically do, while children with autism whose verbal mental ages are higher than 4 years typically do not while matched children with moderate mental handicap (MLD) do (Russell et al., 1991). This study also showed that success on the windows task was related to success on the Wimmer and Perner’s (1983) unexpected transfer task in which children must predict which location a protagonist will visit when he holds a false belief about the current location of a desired object. While this implied some commonality between the windows task and a widely used test for a ‘theory of mind,’ subsequent work has suggested that in some instances, children as young as 3.5 years may pass the windows task (Moore, Barresi, & Thompson, 1998; Russell, Jarrold, & Potel, 1994; see also Samuels, Brooks, & Frye, 1996), making it easier than the false belief task.

As sketched above, although the task was originally designed as a test of strategic deception it also makes unambiguous executive demands, insofar as the child must (1) inhibit the prepotent response of pointing to the desired object while (2) hold in mind the rule that the empty box must be indicated. Moreover, there are two features of younger children’s failure on the task which suggest that they are challenged by it, at least in part, because of the executive demands it makes. In the first place, many younger children continue to make the incorrect response of pointing to the baited box (thereby losing the treat) for as many as 20 test trials. This can be regarded as an example of the kind of ‘stuck-in-set’ perseveration found in frontal patients (Sandson & Albert, 1984). Second, removing the opponent does not remove the difficulty for either typically-developing children (Russell et al., 1994) or children with autism (Hughes & Russell, 1993). That is to say, when the child has to direct the experimenter to the empty box when she is sitting beside the child with the windows visible to him or her, performance is still poor and a high proportion of 3-year-olds perseverate with the wrong response for all 20 trials. In this case the task involves misleading but not deceiving.

In a recent study, Hala and Russell (2001) set out to reduce the inhibitory requirements of the task by having pre-school children indicate their choice of box using an artificial response medium (placing a cardboard pointing hand or a cardboard star on the box). It was thought that the act of pointing in the standard procedure might have proved difficult to inhibit because of its very naturalness. The use of an artificial response medium did indeed lead to improvements in performance but only in the ease with which the correct response was learned during the course of the 20 test trials. It did not encourage a substantially greater number of correct responses on the first, crucial windows trial in which children had to draw an inference about which box to indicate. The manipulation helped 3-year-olds to pull back from perseveration, but it did not help them to infer the utility of the deceptive act, which is what the task had been originally designed to measure.

A substantial residual challenge remains, then, in the windows task when the inhibitory demands are reduced by the use of artificial response media — a
challenge, which we need to identify. In the first place, the task continues to require children to mislead an opponent even when the response is mediated artificially. Indeed it would be argued by those who regard the task as tapping mentalising abilities that these social demands are the crucial ones. Given this possibility, it is necessary to present the task in such a way that other people are not involved at all — either as opponents to deceive (Russell et al., 1991) or as experimenters with whom to compete for resources (Russell et al., 1994). For this reason, we presented both preschool children and two clinical groups (children with autism and with MLD) with a fully automated version of the task in which the participants ‘play against’ a machine rather than a person (see Fig. 1).

However, if automating the task did indeed remove the challenge it is nonetheless possible that this outcome was due to the mechanised nature of the task rather than to the absence a human opponent. For this reason, we also included conditions in which, while the mechanised procedure was used, there was an opponent to play against, either deceptively or non-deceptively. In the opponent-deceptive case the opponent sits on the other side of the machine from the child to ensure that she cannot see the location of the treat in the box, with a correct response being for the child to indicate the button beneath the empty chamber for the opponent to press. In the opponent-nondeceptive condition the opponent sits beside the child and is therefore able to see both locations. In the latter condition, the child has learned that if the button beneath the baited chamber is pressed the sweet is given to the opponent, so the child’s task is to press the button beneath the empty chamber,
after which the sweet is given to him or her. If the challenge presented by the windows task is an essentially social one then both opponent conditions — especially the deceptive one — will continue to challenge young children and children with autism despite its mechanised nature.

It will be proposed that three possible outcomes can highlight the locus of difficulty in the windows task.

1. If the group differences are removed when the child ‘plays against the machine’ while remaining in the two ‘social’ conditions then it would suggest that the locus of difficulty is with whatever forms of mentalising are required for deceiving and competing. This would result in interactions between groups and conditions.

2. If mechanisation of the task has no effect on group differences, it would suggest that the locus of difficulty is executive in nature.

3. Group differences being removed in all three conditions would be consistent with there being something peculiar to the mechanised procedure itself that removed the task’s difficulty; although the nature of this ‘something’ would remain to be determined.

Finally, we will be using two performance criteria: (1) the ability to infer the utility of the correct response on the first ‘windows’ trial; (2) the number of errors made across all 20 trials. As discussed, the study by Hala and Russell (2001) has already demonstrated that using a ‘symbolic’ response medium removes group difference in normally developing children on criterion (2). It will therefore be of particular interest to observe the effect of automating the task on criterion (1) — and in children with autism.

2. Study 1

2.1. Method

2.1.1. Participants

Children were recruited through day-care centres and preschool playgroups in Cambridge, UK. A total of 140 3- and 4-year-old children participated. Data are reported here for the 108 participants who successfully completed the training phase of the study. Three-year-olds: mean age, 3 years 5 months; range, 3:0–3:11; 19 males, 35 females. Four-year-olds: mean age, 4 years 4 months; range, 4:0–4:6; 25 males, 25 females. The excluded children were predominantly younger (mean age \( = 3:5 \), range, 2:11–4:4), 18 did not reach criterion in the training phase, an additional five spoke English as a second language, while nine refused to complete all test trials. The remaining children were assigned randomly to one of three experimental conditions to be described below; 18 children per group. A one-way ANOVA confirmed that the mean ages of the children in the different groups at each age level did not differ.
2.1.2. Design

There were two between-subject variables: task condition (no-opponent, opponent-nondeceptive, opponent-deceptive) and age (3 or 4 years).

2.1.3. Apparatus and materials

A novel automated version of the standard windows task originally designed by Russell and his colleagues (Russell et al., 1991) was introduced (see Fig. 1). The apparatus was 48 cm wide, 26 cm deep, and 24 cm high. Visible to the children were two opaque, cylindrical containers (one red, one blue) sitting on top of the metal platform housing the mechanism. The front face of each of these cylinders displayed an aperture that could be opened surreptitiously by the experimenter using a remote control mechanism the size of a key fob. When opened, these apertures created a transparent window into each cylinder. When the windows were opened, a sweet (a green M&M) was revealed to be inside one of the cylinders. The opening of the windows was accomplished by activating the retraction of opaque sleeves inside of each of the cylinders. Only a single sweet was visible on each trial, located in either the red or blue cylinder. For practical reasons, however, and unbeknown to participants, both cylinders contained permanently mounted identical sweets at all times, though the machine was designed with rotating platforms in order to permit only a single sweet to be viewed at a time. To achieve this effect the inner portion of each cylinder was divided by an opaque white plastic wall, on one side of which there was a green M&M. When the windows were opened this inner section swivelled round. It did so in such a way that the walls moved to 180° of the child and so that the sweet in one container was revealed while the sweet in the other was hidden. The actual M&Ms that were dispensed were contained out of view inside the body of the machine.

2.1.4. Procedure

Participants were seen individually by either one female experimenter (no-opponent condition) or by two female experimenters (both opponent conditions). The first experimenter remained with the child throughout the procedure while the second acted as the “opponent.” Testing took place in the nursery schools, either in a separate room or an area screened off for testing.

All conditions included a training phase prior to the testing phase. The training phase was included to ensure that children understood both the rules of the game and the consequences of their responses. In the training trials the “windows” of the cylinders were closed until after the participants responded with their choice, after which the windows would open to reveal the location of the sweet. The exact procedures of the training phase were matched to the procedures for the corresponding test phase across the three experimental conditions. The training phase for the no-opponent condition will be described in detail. The full verbal protocols for the three conditions are given in Appendix A.
2.1.5. No-opponent condition

The experimenter sat beside the child, both facing the front of the machine. During the introduction to the game the windows remained closed. Children were shown the apparatus and told that they were to play a game “with the machine.” Their attention was drawn to the containers on the top of the apparatus and they were told that there was always a “Smartie” (a term more familiar to English children than M&Ms) in one of them and that their goal was to try to win the Smartie.

2.1.6. Training phase

The experimenter pointed out the buttons beneath the two containers. Children were told that they were to press one of these buttons on each turn. The experimenter explained that, after a button was pressed, the windows on the containers would open revealing the location of the Smartie. Children were then told the rule that if they had pressed the button under the container that held the sweet, they did not receive the sweet. If, however, they had pressed the button under the empty container, then they won the sweet.

After these initial instructions, children were told to press one of the buttons when the experimenter said “ready.” Immediately after they had pressed a button, both windows opened to reveal the location of the Smartie. Once the child had seen which container held the sweet the experimenter emphasised the rule by providing explicit feedback. If the child happened to have pressed the button beneath the baited container the experimenter announced: “Look, the Smartie is in the one you pressed the button for. That means no Smartie for you this time”. The windows then closed and the next trial began. Alternatively, if the child happened to press the button under the empty container the experimenter announced: “Look, the Smartie is in the other one — the one you didn’t press. That means you get to have the Smartie.” The windows then closed and the machine dispensed the sweet.

The explicit feedback continued for four trials. On the fifth trial, after the windows had opened to reveal the whereabouts of the sweet, the experimenter asked one of the following questions: “There’s a Smartie in the one you pressed, so do you get the Smartie or does it stay in the machine?” or “There’s no Smartie in the one you pressed, so do you get the Smartie or does it stay in the machine?” When the child had answered correctly on three consecutive trials the training phase was terminated and the test phase began. If this criterion was not reached within 15 trials the child did not progress to the test phase.

2.1.7. Test phase

Before the test phase began, the experimenter told the child that she had to do something to the machine. She then changed the sweet-dispensing carousel by opening the machine at the back, reloaded it, and reset the machine via a switch at the back. During this time the windows remained shut. The child was told that it was the same game and that the rules were the same as before. The windows were opened and the experimenter said, “Look, this time the windows are open already. Now you can see which one has the Smartie, so now it will be easier
to win Smarties. Okay, press one of the buttons.” If the child pressed the button beneath the empty container a sweet was dispensed immediately after the windows closed (as if the visible sweet was being dispensed), if the child pressed the button beneath the baited container, the windows closed but no sweet was dispensed (as if the sweet was being returned to the machine). No explicit feedback regarding the rule was given during the test trials. The experimenter simply announced the result in a neutral tone saying either: “The Smartie stays in the machine. Let’s try again.” or “You get the Smartie this time. Let’s try again.” This continued for 20 trials.

2.1.8. Opponent-nondeceptive condition

In this condition, children were seen by two experimenters. The experimenters sat on either side of the child, all three facing the front of the machine. The second experimenter acted as the opponent. Children were told that they would play a game with the opponent (named) to try to win sweets from her. Children were introduced to the rule: “If the Smartie is in the one you pressed, that means you don’t get to have it and (opponent) gets it. If it’s in the other one though, you get it.” The general procedures and protocols were otherwise the same as in the no-opponent condition except that children were now competing with an opponent rather than a machine (see Appendix A for full details of the verbal protocol). In this condition, as in the no-opponent condition, it was always the child who selected and pressed the button. In contrast to the no-opponent condition, however, in the opponent-nondeceptive condition a sweet was dispensed irrespective of the button that the child had pressed. The first experimenter then awarded the sweet to the appropriate player. Specifically, if the child had pressed the button under the baited container, the sweet was awarded to the opponent. Alternatively, if the child had pressed the button under the empty container the sweet was awarded to the child.

As in the no-opponent condition, the children received a training phase containing explicit feedback about the contingencies for four trials and from the fifth trial onwards they were asked to say who had won the sweet. The criterion for continuing to the test phase was the same as in the no-opponent condition. The procedure and protocols of the test phase were modified to take account of the fact that the competition was against another person rather than against the machine.

2.1.9. Opponent-deceptive condition

This condition reproduced the general procedure of the standard version of the windows task (Russell et al., 1991) as far as possible within the constraints introduced by the machine. Accordingly, in this condition the second experimenter (opponent) sat across the table from the child and the first experimenter. Consequently the opponent was seated behind the automated windows machine, facing the back of the machine and was thus unable to see into the containers when the windows were open. As with the other two conditions, the first experimenter sat beside the child, both facing the front of the machine. As in the opponent-nondeceptive condition, children were told that they would be playing a game with the opponent to try to win Smarties from her.
In the original windows task, the successful strategy was one that misled the opponent to open the empty box. Specifically, if children directed their opponent to open the empty box, they won the sweet. In our automated procedures, however, the cylinders were sealed so we needed to employ an alternative avenue that would also allow children to misdirect their opponent. We accomplished this by having children select the button that the opponent was to press. Children were instructed to point to, but not to press, the button on one of the two containers when the experimenter said “ready.” The instructions took the following form: “When I say ready you point to the button that you want (opponent) to press... If the Smartie is in the one (opponent) presses the button for, that means she gets to keep it. If it’s in the other one, then you get to keep the Smartie.” In this condition, in contrast to the previous two conditions, while the child selected and pointed to the button it was always the opponent who pressed it. In the training phase, as in the other conditions, neither the child nor the opponent saw the location of the sweet before the opponent pressed the button. As in the opponent-nondeceptive condition, a sweet was dispensed on every trial, with the first experimenter allocating it to the appropriate player according to the rule.

As in the other two conditions, explicit feedback was provided for the first four training trials. Again, in subsequent trials the child was asked who should receive the sweet when it was dispensed, and again the child had to give the correct answer in three consecutive trials as a criterion for continuing to the test phase.

In the test phase, since the windows faced the child, the child, but not the opponent, could see the location of the sweet when the windows were open. The child was instructed to point to the button while the windows were open. After the child had indicated which button the opponent was to press, the windows closed (via remote control) and the opponent came round to the front of the box and pressed the button that had been indicated by the child. Otherwise the procedure was the same as for the opponent-nondeceptive condition.

2.2. Results and discussion

The main question addressed in this study was whether automating the procedure would impact upon the typically poor performance of 3-year-olds on the windows task. Reported here are comparisons of the number of children in both age groups who made correct responses on the first trial, performance across all trials, and the incidence of better-than-chance performance across trials. See Table 1.

2.2.1. Performance on trial one

It is clear from the table that there was little difference between the first trial performance of the two age groups. Logistic regression analyses comparing performance on trial one between age groups and across conditions confirmed that there was no significant effect of age. This finding contrasts with that of studies employing standard versions of the windows task (e.g., Hala & Russell, 2001; Russell
Table 1

Performance of normally-developing children on the three conditions of the automated windows task

<table>
<thead>
<tr>
<th>Condition</th>
<th>No-opponent (n = 18)</th>
<th>Opponent-nondeceptive (n = 18)</th>
<th>Opponent-deceptive (n = 18)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First trial correct</td>
<td>13</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>Mean errors per trial</td>
<td>2.94</td>
<td>3.35</td>
<td>6.65</td>
</tr>
<tr>
<td>Perseverate on all trials</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First trial correct</td>
<td>14</td>
<td>14</td>
<td>10</td>
</tr>
<tr>
<td>Mean errors per trial</td>
<td>1.85</td>
<td>1.30</td>
<td>2.95</td>
</tr>
<tr>
<td>Perseverate on all trials</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

et al., 1991; Russell et al., 1994), wherein 3-year-olds were found to perform significantly worse on trial one than 4-year-olds.

An alternative way to evaluate the effectiveness of the automated procedure is to examine whether performance differed from chance on the first trial. For the 3-year-olds, first trial performance was indeed better than chance for the no-opponent condition (72% correct, $P < .05$, Binomial test) but not different from chance on the other two conditions (56% for the opponent-nondeceptive; 61% for the opponent-deceptive; $P > .1$). For the 4-year-olds, first trial performance was better than chance for the no-opponent and the opponent-nondeceptive conditions (78% correct for both, $P < .05$) but not for the opponent-deceptive condition (56% correct, $P > .1$).

Of particular interest here is the performance of the 3-year-olds relative to the typical pattern of failure on the first trial when the procedure is not automated. While the first trial performance of the 3-year-olds in the two opponent conditions reported here was not significantly above chance, it was superior to that of the same-aged children in standard versions of the windows task (e.g., Russell et al., 1991, 1994). Indeed this was also true in the case of the 3-year-olds whose performance across trials had benefited from the symbolically-mediated response procedures used by Hala and Russell (2001) in a non-automated version of the task. Although cross-study comparisons must be interpreted with caution, it is nonetheless useful to compare the first trial performance of the 3-year-olds in the two ‘opponent’ groups (36 children, 21 correct) with the first trial performance of the 3-year-olds in the non-automated symbolic-media groups reported in Hala and Russell’s Study 3 (45 children, 16 correct). The performance of the children in the automated windows task with an opponent was significantly superior ($\chi^2(2) = 4.09$, $P < .05$). This comparison suggests that while the first trial performance of the two opponent conditions in the current study did not exceed chance, automating the procedure brought about better first trial performance than did providing artificial response media.1

1 The 3 year olds in this automated version of the study, as well as being of similar ages to those in the non-automated version of the task, were also from the same or similar schools.
2.2.2. Performance across trials

Performance throughout the 20 test trials was compared across task conditions and age groups. An age (3/4 years) by condition (no-opponent, opponent-nondeceptive, opponent-deceptive) ANOVA was carried out with total score correct across trials as the dependent variable. There were significant main effects of both age \( F(1, 102) = 8.24, P < .01 \) and task \( F(2, 102) = 3.74, P < .05 \), but no significant interaction between these factors.

Figs. 2 and 3 highlight the effect of age. The overall superior performance of the 4-year-olds appears to be due to the fact that they were more likely to maintain adequate performance throughout the 20 trials. In contrast to previous studies, this superior performance was not due to the fact that older children were better on the initial trials, as we have seen. The effect of task was examined with post hoc Tukey tests. None of these individual comparisons was significant.

2.2.3. Above-chance responding across trials

When logistic regression analysis was conducted on the numbers of children who performed at a level better than chance across the 20 trials (i.e., 15 or more trials correct) there was found to be an effect of age \( \text{Wald}(1) = 4.33, P < .05 \). There was, however, no effect of condition across age or within each age. Inspection of Figs. 2 and 3 shows that these analyses must, nevertheless, be treated with some caution, at least as regards the difference between the opponent-deceptive condition and other two conditions at age 3. It can be seen from Fig. 2 that, after the first trial, performance on the opponent-deceptive condition is uniformly worse than on the other two conditions for the subsequent 19 trials, apart from on trial 9 when it is equal to that on the no-opponent condition.

Our next study employed exactly the same procedures as Study 1, but with children with autism and with moderate learning difficulty as the comparison groups. As set out in Section 1, the difficulties with the windows task experienced by children with autism are well established, so we wished to determine whether automating the procedure would have an effect similar to that produced in our typically-developing preschool children.

3. Study 2

3.1. Method

3.1.1. Participants

A total of 128 children with autism or MLD participated in the second study. Data are reported for the 117 participants who successfully completed the training phase of the study. In the opponent-nondeceptive condition one child with autism (verbal mental age 2.05) and two children with MLD (verbal mental age 4.06 and 5.02) did not progress beyond the training phase. In the opponent-deceptive condition, five children with autism (verbal mental age between 3.00 and 6.02)
Fig. 2. The number of children in each of the three task conditions indicating the correct window in each test trial at age 3.
Fig. 3. The number of children in each of the three task conditions indicating the correct window in each test trial at age 4.
Table 2
Participant details for Experiment 2

<table>
<thead>
<tr>
<th></th>
<th>CA</th>
<th>VMA</th>
</tr>
</thead>
<tbody>
<tr>
<td>No-opponent (I)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism (n = 20)</td>
<td>12.58 (2.39)</td>
<td>5.30 (2.35)</td>
</tr>
<tr>
<td>MLD (n = 20)</td>
<td>10.33 (2.03)</td>
<td>5.40 (2.49)</td>
</tr>
<tr>
<td>Opponent-nondeceptive (II)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism (n = 20)</td>
<td>8.35 (1.48)</td>
<td>5.25 (1.89)</td>
</tr>
<tr>
<td>MLD (n = 20)</td>
<td>9.40 (1.21)</td>
<td>5.95 (1.62)</td>
</tr>
<tr>
<td>Opponent-deceptive (III)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autism (n = 20)</td>
<td>8.90 (2.92)</td>
<td>4.91 (2.20)</td>
</tr>
<tr>
<td>MLD (n = 17)</td>
<td>9.75 (1.04)</td>
<td>5.35 (1.64)</td>
</tr>
</tbody>
</table>

Ages are shown in years (standard deviations).

and two children with MLD (verbal mental age of 3.06 and 4.01) did not progress beyond the training phase.

Children were allocated to one of the three conditions of the automated windows task described in the previous study in order to ensure that the groups of children in any one of the three task conditions did not differ from any other on the basis of verbal mental age, as measured by the British Picture Vocabulary Scale (BPVS-Long Form; Dunn, Whetton, & Pintillie, 1982). A one-way ANOVA with verbal mental age as the dependent variable showed that there was no difference between the groups on this measure \( F(1, 116) = 1.25, P > .1 \). The gender ratios (male:female) were 9:1 for the children with autism and 1.59:1 for the children with MLD. Participant details are shown in Table 2.

The children with autism were selected if they had been formally diagnosed as having autism. All children in this group were attending special schools or units for children with autism in the UK. The children with MLD were attending schools for children with learning difficulties in the UK and were included only if they did not have a specific diagnosis such as Down syndrome.

3.1.2. Design

There were two between-subject variables: task condition (no-opponent, opponent-nondeceptive, and opponent-deceptive) and participant group (autism, MLD).

3.1.3. Procedure

The task was undertaken exactly as described in the previous study, with children being tested individually in a quiet room in their school.

3.2. Results and discussion

Corresponding to Study 1, the main question addressed in this second study was whether using an automated version of the windows task would ameliorate
Table 3
Performance of children in the two clinical groups on the three conditions of the automated windows task

<table>
<thead>
<tr>
<th></th>
<th>No-opponent (n = 20)</th>
<th>Opponent-nondeceptive (n = 20)</th>
<th>Opponent-deceptive (n = 20)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Autism</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First trial correct</td>
<td>15</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>Mean errors per trial</td>
<td>3.45</td>
<td>3.6</td>
<td>6.70</td>
</tr>
<tr>
<td>Perseverate on all trials</td>
<td>0</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>MLD</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First trial correct</td>
<td>15</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td>Mean errors per trial</td>
<td>2.55</td>
<td>1.00</td>
<td>5.40</td>
</tr>
<tr>
<td>Perseverate on all trials</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

the typically poor performance of children with autism. The performances of the children with autism and MLD are shown in Table 3.

3.2.1. Performance on trial one
We asked whether automating the procedure would bring about generally successful performance on the first test trial. It is clear from these tables that there was little difference between the first trial performance of the two participant groups. Logistic regression analyses comparing performance on trial one between the two groups and across task conditions confirmed that the children with autism did not differ significantly from the children with MLD. Although no main effect was found for group there was a significant effect for condition [Wald(2) = 6.35; P < .05]. Paired contrasts indicated that performance in the opponent-nondeceptive condition was significantly lower than in the no-opponent condition [Wald(1) = 6.26, P < .05]. No other contrasts were significant.

In comparing first trial performance against chance levels, both groups were better than chance in the no-opponent condition (75% correct: P < .05, Binomial test). Performance was not significantly different than chance on the other two conditions (P > .1 on Binomial). Thus, the overall pattern of performance on trial one was similar across the two groups.

3.2.2. Performance across trials
Performance throughout the 20 test trials was compared across task conditions and participant groups. A group (autism, MLD) by task condition (no-opponent, opponent-nondeceptive, opponent-deceptive) ANOVA was carried out with total score correct across trials as the dependent variable. There was a significant main effect of participant group [F(1, 111) = 6.88, P < .01] with the children with MLD demonstrating superior performance overall. There was no significant main effect of task [F(2, 111) = 2.66, P > .1], nor was there a significant interaction between group and task condition [F(2, 111) = .97, P > .1]. An examination of Figs. 4 and 5 illustrates that the overall higher performance of the children with MLD primarily arises from their strong performance in the opponent-nondeceptive...
Fig. 4. The number of children with autism in each of the three task conditions indicating the correct window in each test trial.
Fig. 5. The number of children with MLD in each of three experimental conditions indicating the correct window in each test trial.
condition. Independent $t$-test analyses indicated a significant difference between the two groups only for this condition [no-opponent, $t(38) = -0.63, P > .1$; opponent-nondeceptive, $t(38) = -2.52, P < .05$; opponent-deceptive, $t(38) = -1.82, P > .1$].

3.2.3. Above-chance responding across trials

We conducted logistic regression analyses comparing the proportion of children who performed at above chance levels (i.e., 15 or more trials correct) both between the two groups and across the three experimental conditions. These analyses yielded a main effect for group [Wald(1) = 8.25, $P < .01$]. No significant effects were found for task, nor was there a significant interaction between group and task. These results should, however, be treated with caution, as closer examination of Figs. 4 and 5, shows that the pattern of responses for both groups on the opponent-deceptive condition was consistently below that of their performance on the other two task conditions.

4. General discussion

We consider, first of all, how these data match the three possible, theoretically informative outcomes listed in the Introduction. These were: (1) automisation of the task making the task easy to pass in the no-opponent condition but not in the two opponent conditions; (2) automatisation having no effect, with the substantial challenge to typically-developing 3-year-olds and to children with autism remaining; (3) automisation essentially removing the challenge of the task and therefore removing the group differences that are usually found. These would each imply the following about the locus of task difficulty: (1) the nature of the challenge is essentially a social/mentalising one; (2) the challenge is essentially executive in nature; (3) something peculiar to the mechanisation of the task is responsible for the removal of the challenge. We will argue that, while these two studies provided partial evidence for all three of these outcomes, (3) received the strongest support.

In the first place (1) the no-opponent condition was indeed the easiest and the opponent-deceptive condition was the most difficult condition. In fact, when only responses to the first ‘inferential’ trial are taken into consideration the participants overall were significantly better on the no-opponent condition than on the other conditions. This provides some support for the idea that the essential challenge of the task is social. Set beside this, however, is the fact that performance on the two opponent conditions was clearly superior to that usually found with the non-automated procedure; which will be discussed more fully below. Most strikingly, perseveration across all trials was very rare, in contrast to the fact that the majority of 3-year-olds and children with autism typically continue to point to the baited box over 20 trials when the procedure is not automated (e.g., Hala & Russell, 2001; Hughes & Russell, 1993; Russell et al., 1991, 1994). Additionally, we did not find the predicted interactions between groups and conditions.
With regard to the second possibility (no effect of automisation), the positive evidence here is the fact that the significant superiority of the 3-year-olds over the 4-year-olds and of the children with MLD over the children with autism remained — at least on the criterion of performance across all trials. This might be said to support the view that the essential challenge of the task is executive *tout court*; though see below for a more subtle version of the executive position. But set against this, as we have seen, is the fact that performance was far better overall than in the standard, non-automated procedure. Performance was better than chance, or approaching this, in many sub-groups on the first inferential trial and the amount of perseveration was extremely low in all conditions.

The evidence in favour of the third possibility (automisation essentially removing the challenge to the 3-year-olds and children with autism) is more substantial. As has been discussed, first-trial performance in these children was frequently good and perseveration was rare among them. While first-trial performance was no better than chance in the two opponent groups, it did not lag far behind in the case of the 3-year-olds; see Table 1. Indeed, recall that the first-trial performance of the typically-developing children in the automated task with an opponent was significantly superior to that found by Hala and Russell (2001) when a symbolic response medium was used ($\chi^2(2) = 4.09, P < .05$). It has to be said, however, that the group effects that continued to be found on the criterion of performance across all trials does weaken this conclusion.

This somewhat anomalous set of results might be made to yield a clearer message by addressing the issue of whether first trial performance or performance across all 20 trials is the more sensitive measure. If a case can be made for first-trial performance being a good index of the task’s essential challenge then the third possible outcome (challenge removed) could be said to be the more strongly supported. With regard to the opposite claim (across-trial performance more informative), a study by Gerstadt, Hong, and Diamond (1994), using a measure of prepotent response inhibition in typically-developing pre-school children, showed that the younger children’s difficulty was with *maintaining* initially adequate performance across 16 trials. This fact might be used to suggest that, if it is the executive nature of the challenge that is at issue, then the more sensitive measure is performance across trials, not performance on the first, inferential trial. However the Stroop-like ‘day–night’ task used by these authors, in which children have to say “day” to a night sky and vice versa, is conceptually distinct from the windows task, in the following respect. The children in Gerstadt et al’s study were told explicitly what they had to do, whereas in the windows task participants have to work this out for themselves on the first trial (and utilise negative feedback on subsequent trials). The windows task can therefore be regarded as a *measure of rational action when that action is counter-intuitive*; it is not just a test of following counter-intuitive instructions — not a kind of Stroop test. This is what might account for the fact that children with autism do not show a deficit on the day–night task (Russell, Jarrold, & Hood, 1999). Given this, it can be argued that it is first-trial performance that captures what is distinctive to the task. Recall, also, that Hala and
Russell (2001) found that across-trials group differences among pre-school children could be removed by using of a symbolic response medium, while first-trial group differences remained. It would appear that the removal of first-trial group difference with automisation is what requires explanation.

This leads naturally to the question of how the improvement with automisation came about. What challenge did automisation remove? The two main differences between the standard procedure and the automated one were in respect of the delivery of the reward and in respect of the mode of response. As it is the latter that is a direct measure of what children have to do, this would appear to be the crucial difference. In the no-opponent and in the opponent-nondeceptive conditions the participants had to press a button rather than indicate a certain location (as in the standard procedure); and these two conditions were the most successful. The mode of response required in the opponent-deceptive condition, however, was a hybrid of pressing and indicating a location — indicating which button had to be pressed. It could be argued, then, that the fact that this condition required a response that was, to some degree, similar to that used in the standard task may have accounted for its continuing to present something of a challenge.

The implications of this analysis are the following. When children have to bring about a desired effect mechanically, in either a competitive or a non-competitive context, they cease to be challenged by a task that has all the hallmarks (executive and social) of a 3–4 transition task and one on which children with autism show a specific deficit. The context in which an explanation of this result will be offered is that of two different kinds of inhibition. In a nutshell, 3–4 transition tasks (ones that also challenge individuals with autism) require a form of inhibition that the automated windows task does not, namely, inhibiting reference to a salient state of affairs, as opposed to inhibiting a means-end action near a goal object (required by the present task). The discussion will initially concern typically-developing children, later turning to children with autism.

This kind of account of developmental transition by reference to inhibitory demands is essentially nativist in character. The claim is that children under 4 years essentially possess a theory of mind, but that it is one with very limited application partly due to their executive immaturity. Accordingly, there is seen to be no discontinuity between 3- and 4-year-old ‘mentalising.’ The following analogy with an issue in syntactic development may be useful here. Ninio (1988), contrary to Braine (1976) and to Bowerman (1976), argued that the fact that young children’s two-word combinations may be analysable as ‘limited-scope formulae’ does not negate the possibility that these productions are, in fact, restricted versions of adult grammar. Accordingly, a child whose use of the verb want may be parsimoniously described in terms of the formula ‘want + X,’ rather than as ‘verb + noun,’ can nevertheless be credited with the symbol-string ‘VP → V + NP’ but with strict limitations on the semantic character of the verb (V ← want). To do otherwise is to leave ourselves with the challenge of explaining the radical transition between the use of limited-scope formulae and full competence.
Similarly, one might say that the under-4’s conceptions of belief, intention, and so forth, are not qualitatively different from those of older children, but that when the younger children are asked questions with a certain executive structure they fail to inhibit reference to salient states of affairs. Thus, we explain their failure on a range of tasks that are normally failed at age 3 and passed at age 4 in terms of failure of inhibition of reference to these salient states of affairs (Russell, 1996; Russell et al., 2001). The most notable among these tasks are the unexpected transfer task (Wimmer & Perner, 1983), unexpected contents task (Perner, Leekam, & Wimmer, 1987), and the appearance-reality task (Flavell, Flavell, & Greene, 1983). In these cases, the child is aware of something in the moment of questioning that blocks the framing of the correct answer, be it the current location of the focal object, the actual nature or contents of a deceptive object, or the phenomenal result of a property change. Failure takes the form of failing to inhibit reference to a salient fact about reality or appearance.

The current claim is, then, that the standard version of the windows task also requires this kind of inhibition, but that the automated version (most clearly in the no-opponent and the opponent-nondeceptive conditions) does not. The implication is not simply that young children do not have a generalised difficulty with the inhibition of inappropriate means-end action, but that the ability to engage in the kind of strategy tapped by the windows task is present in young children. By extension, one could also say that the data are consistent with related abilities being present more generally in the domain of unequivocal theory of mind tasks.

With regard to children with autism, data from another study suggest that these individuals do have substantial difficulty with inhibiting inappropriate means-end actions at mental ages at which typically-developing do not. In Hughes and Russell’s (1993) detour-reaching task, in which children had to throw a switch to enable a direct reach into a box for a reward, individuals with autism were found to be specifically challenged; indeed some participants with verbal mental ages of 10 or 11 years were found to fail to reach criterion. By contrast, many typically-developing 2-year-olds attain criterion on this task (Biró, 2001). This of course, raises the question of why the children with autism performed so well in the present study. This might have been because, in contrast to the response required on the detour-reaching task, the action to be taken was not a highly practiced one in the present case.

To return finally to typically-developing children, the main competitor to the view that development of the executive system around age 3 or 4 years enables children’s mentalising abilities to come to fruition is Perner’s view that acquiring the conception of an intentional mental state’s power to cause action makes inhibition possible (Perner & Lang, 1999). On the latter view, children’s understanding of how a protagonist’s false belief causes him to visit the wrong box (Wimmer & Perner, 1983) finds a parallel in their appreciation that seeing the experimenter’s hand-shape in the Luria hand game causes them to produce the wrong response of matching their hand to her hand. It is notable in the light of the present data that
Perner and Lang (1999) report that the children’s version of the Luria hand game correlates much less substantially with false belief performance than does a task in which children must judge a knee-reflex evoked from them not to have been intended. On Perner’s position, this is because the knee-reflex task is a direct test of a theoretical insight into the causal power of intentions, whereas the hand-game is just another executive task whose successful completion manifests this insight. On the present view, however, both tasks are executive tasks, with the knee-reflex task requiring the inhibition of an answer based on perceived outcome (see Russell et al., 2001). On the present view, the reason why it correlates so strongly with the false belief task is that the response that has to be inhibited is a (verbal) referential act, in contrast to the motor inhibition necessary in the Luria hand game (cf. the Hughes & Russell, 1993, detour-task, mentioned above).

Finally, we make a prediction based on the position articulated here. If it is indeed referential acts and not actions that 3-year-olds find difficult to inhibit, then we should see them performing adequately on a ‘false action’ task. Consider a situation in which a child sees a protagonist being shown that one kind of action brings about a desired result (e.g., playing music), after which the apparatus is re-engineered, in his absence, so that this action ceases to be effective but a new action has become effective. The false-action question (asked about the protagonist on his return) is: ‘Show me what X will do to make the music play?’ General success on such a task by 3-year-old children would reinforce the claim that when reference to salient a states of affairs does not have to be inhibited, young children can evince knowledge of how a false assumptions can lead to erroneous behaviour.

Acknowledgments

This research was made possible by grants from the Economic and Social Research Council (UK) to J. Russell and S. Hala and from the Wellcome Trust (UK) to J. Russell. For participation in Study 1 we would like to express appreciation to the following nursery schools and play-centers in Cambridge, UK: Ace, Anglia Polytechnic, Harvey Road, Joint Colleges, King’s Street, Playlander’s, Richmond Road, Ross Street, Rockinghorse and Studio. We are grateful to the headteachers, staff and pupils of the following schools for their kind cooperation and support with the research reported in Study 2: Ashgrove House, Penarth; Axton Chase School, Longfield; Lady Adrian School, Cambridge; C. Langafel of E. School, Longfield; Marshfields School, Peterborough; Newick House School, Burgess Hill; Springfields School, Huntingdon; Walnuts School, Milton Keynes; Wargrave House School, Newton-Le-Willows; Woolgrove School, Letchworth. We appreciate the help of Emma Clubley, Rebecca Saltmarsh and Melanie Senior with a portion of the data collection. Thanks also go to Andreas Mehne for his drawing of the machine. Finally, we extend our gratitude to Ray Rowles of Mars Confectionary who so generously kept us supplied with green M&Ms.
### Appendix A. Testing scripts across condition

#### 4.1. Training phase

<table>
<thead>
<tr>
<th>Condition</th>
<th>No-opponent</th>
<th>Opponent-nondeceptive</th>
<th>Opponent-deceptive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Introduction</strong></td>
<td>Windows closed. E1 sits next to child. “I have a game for you to play. It’s a game that you play by yourself with this machine. You see these two things on the top here? (Points to cylinders). There’s a Smartie inside one of them. You’re going to play a game to see if you can win that Smartie”</td>
<td>Introduction. Windows closed. E1 and E2 (opponent) sit beside child, with child between them. “I have a game for you to play with E2. It’s a game that you play with this machine. You see these two things on the top here? There’s a Smartie inside one of them. You’re going to play a game to see if you can win that Smartie from E2”</td>
<td>Introduction. Windows closed. E1 sits beside child, E2 (opponent) sits at back of machine. “I have a game for you to play with E2. It’s a game that you play with this machine. You see these two things on the top here? There’s a Smartie inside one of them. You’re going to play a game to see if you can win that Smartie from E2”</td>
</tr>
<tr>
<td><strong>Training phase</strong></td>
<td>“See these two buttons down here? When I say ready you press one of these buttons. When you do that these little windows will open and you will see which container has the Smartie. If the Smartie is in the one you pressed, you don’t get to have it. If it’s in the other one though, you do get it. Ok, ready, press a button.” Child presses button — windows open</td>
<td>Training phase. “See these two buttons down here? When I say ready you press one of these buttons. When you do that these little windows will open and you will see which container has the Smartie. If the Smartie is in the one you pressed, you don’t get to have it and E2 gets it. If it’s in the other one though, you get it.” Ok, ready, press a button.” Child presses button — windows open</td>
<td>Training phase. “See these two buttons down here? When I say ready you point to the button you want E2 to press. When she presses the button these little windows will open and you will see which container has the Smartie. If the Smartie is in the one E2 pressed the button for, she gets to keep it. If it’s in the other one, then you get it. Ok, ready, point to the one you want E2 to press.” Child points — E2 comes from behind machine and presses button — windows open</td>
</tr>
</tbody>
</table>
### Appendix A. (Continued)

<table>
<thead>
<tr>
<th>No-opponent</th>
<th>Opponent-nondeceptive</th>
<th>Opponent-deceptive</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Feedback.</strong> 1. “Look the Smartie is in the one you pressed the button for. That means no Smartie for you this time”</td>
<td><strong>Feedback.</strong> 1. “Look the Smartie is in the one you pressed the button for — that means no Smartie for you this time. E2 gets it”</td>
<td><strong>Feedback.</strong> 1. “Look the Smartie is in the one E2 pressed the button for. That means E2 gets to keep it and there’s no Smartie for you this time”</td>
</tr>
<tr>
<td>2. “Look the Smartie is in the other one — the one you didn’t press. That means you get to have the Smartie.” Sweet is dispensed from machine when selected container is empty</td>
<td>2. “Look the Smartie is in the other one — the one you didn’t press — that means you get to have the Smartie.” Sweet dispensed every trial — E1 awards it to appropriate player</td>
<td>2. “Look the Smartie is in the other one — the one E2 didn’t press. That means you get to have the Smartie.” Sweet dispensed every trial — E1 awards it to appropriate player</td>
</tr>
<tr>
<td><strong>Criterion.</strong> Trial 5 — windows open but before sweet dispensed: 1. “There’s a Smartie in the one you pressed — do you get that Smartie or does it stay in the machine? “turns — or 15 trials without reaching criterion</td>
<td>2. “There’s no Smartie in the one you pressed — do you get that Smartie or does E2 get it?” Training continues until child correct on 3 consecutive turns</td>
<td>2. “There’s no Smartie in the one E2 pressed — do you get that Smartie or does E2 get it?” Training continues until child correct on 3 consecutive turns — or 15 trials without reaching criterion</td>
</tr>
</tbody>
</table>
Test phase. “Now I have to do something to the machine for a minute.” E1 changes dispenser disk at the back of machine and resets machine — out of sight of participant. Windows remain shut.

“Ok, now we’re ready to start the game again. It’s the same game; the rules are the same as before.” Machine is started on test trials — windows are now open at beginning of each trial.

Windows open. “But look — this time the windows are open already. Now you can see which one has the Smartie — so now it will be easier to win Smarties.”

“Ok — press one of the buttons.” Throughout test phase participants are encouraged to win Smarties but no explicit feedback regarding the consequences of their action is given. Repeat for 20 trials.

Test phase. “Now I have to do something to the machine for a minute.” E1 changes dispenser disk in back of machine and resets machine — out of sight of participant. Windows remain shut.

“Ok, now we’re ready to start the game again. It’s the same game; the rules are the same as before.” Machine is started on test trials — windows are now open at beginning of each trial.

Windows open. “But look — this time the windows are open already. Now you can see which one has the Smartie — so now it will be easier to win Smarties.”

“Ok — press one of the buttons.” Throughout test phase participants are encouraged to win Smarties but no explicit feedback regarding the consequences of their action is given. Repeat for 20 trials.

Test phase. “Now I have to do something to the machine for a minute.” E1 changes dispenser disk in back of machine and resets machine — out of sight of participant. Windows remain shut.

“Ok, now we’re ready to start the game again. It’s the same game; the rules are the same as before.” Machine is started on test trials — windows are now open at beginning of each trial.

Windows open. “But look — this time the windows are open already. Now you can see which one has the Smartie — so now it will be easier to win Smarties.”

“Ok — point to the one you want E2 to press.” (The experimenter closes the windows by remote control when the opponent moves round to be on the same side as the child). Throughout test phase participants are encouraged to win Smarties but no explicit feedback regarding the consequences of their action is given. Repeat for 20 trials.
References


