PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Chimpanzee 'folk physics': bringing failures into focus

Amanda Seed, Eleanor Seddon, Bláthnaid Greene and Josep Call

Phil. Trans. R. Soc. B 2012 **367**, doi: 10.1098/rstb.2012.0222, published 27 August 2012

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

CrossMark click for updates

Phil. Trans. R. Soc. B (2012) **367**, 2743–2752
doi:10.1098/rstb.2012.0222

Research

Chimpanzee 'folk physics': bringing failures into focus

Amanda Seed[1,*], Eleanor Seddon[1], Bláthnaid Greene[1] and Josep Call[2]

[1]*School of Psychology, University of St Andrews, St Mary's Quad, St Andrews KY16 9JP, UK*
[2]*Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, Leipzig 04103, Germany*

Differences between individuals are the raw material from which theories of the evolution and ontogeny of cognition are built. For example, when 4-year-old children pass a test requiring them to communicate the content of another's falsely held belief, while 3-year-olds fail, we know that something must change over the course of the third year of life. In the search for what develops or evolves, the typical route is to probe the extents and limits of successful individuals' ability. Another is to focus on those that failed, and find out what difference or lack prevented them from passing the task. Recent research in developmental psychology has harnessed individual differences to illuminate the cognitive mechanisms that emerge to enable success. We apply this approach to explaining some of the failures made by chimpanzees when using tools to solve problems. Twelve of 16 chimpanzees failed to discriminate between a complete and a broken tool when, after being set down, the ends of the broken one were aligned in front of them. There was a correlation between performance on this *aligned* task and another in which after being set down, the centre of both tools was *covered*, suggesting that the limiting factor was not the representation of connection, but memory or attention. Some chimpanzees that passed the aligned task passed a task in which the location of the broken tool was never visible but had to be inferred.

Keywords: chimpanzees; tool-use; representation; executive function

1. INTRODUCTION

The first step in developing a theory of cognitive evolution or development is to map out the landscape of similarities and differences in cognitive ability. In the case of evolution, this means plotting the cognitive abilities possessed by different species (such as reasoning about objects, time or other minds) on a phylogeny to make inferences about when they are likely to have evolved (using the concept of parsimony), and perhaps even what pressures and evolutionary processes drove their emergence. Unfortunately, this first step is no small hurdle. Measuring cognition is fundamentally different from measurement in many other sciences: first because it is not possible to isolate one compound from the matrix in which it is embedded in nature, unlike chemical processes such as distillation; second and perhaps as a consequence, there are scarcely any 'litmus' tests that are universally accepted as diagnostic of a particular ability. This inherent difficulty has been met with a certain kind of conservatism by psychologists—tests for cognitive abilities such as 'theory-of-mind' or 'causal reasoning' are painstakingly designed to rule out the possibility that simpler mechanisms, such as task-specific learning, could explain successful performance (see [1] for discussion). But when individuals fail the task, should experimenters accept the null hypothesis that the ability is absent? Often it is not that simple, because while the test is designed to narrow down the possible explanations for a positive result, a negative result could have several causes. This is because many tasks probing a particular cognitive skill simultaneously tax other mechanisms. In the developmental literature, some progress has been made in addressing the issue of what causes young children to fail social and physical problem-solving tasks. By briefly reviewing the approaches taken in this area, we hope to extract some general principles that could be applied more broadly.

* Author for correspondence (ams18@st-andrews.ac.uk).

One contribution of 14 to a Discussion Meeting Issue 'Animal minds: from computation to evolution'.

2. WHAT DETERMINES SUCCESS OR FAILURE? CASE STUDIES FROM DEVELOPMENTAL PSYCHOLOGY

(a) *False-belief tasks*

In the developmental literature on the question of 'theory-of-mind', the 'change of location' test designed to probe for belief attribution has presented researchers with considerable interpretive dilemmas. The test requires children to follow a story in which, for example, Anne moves Sally's marble from one hiding place to another out of Sally's sight [2,3]. To pass the task, children have to report, when prompted, that Sally will look for her marble in the place that she saw it last, rather than where it was moved to. Typically developing

children pass this test at around age 4—correctly responding that Sally will look for the marble in its original hiding place [4]. This is consistent with the idea that children at this age have a representational theory-of-mind, and understand that other people can hold 'false beliefs' that differ from what the observer knows to be the case, although there has been some debate about whether a positive result on this test or those like it can support such a conclusion, which is outside the scope of this review [5]. Interpreting the negative result is even more difficult. Do the younger children that fail the task necessarily lack ability to attribute beliefs to others? Even if passing the Sally–Anne task requires (and can therefore provide positive evidence of) theory-of-mind, failing it may be better explained by a lack or deficit in another process that precludes the diagnosis of false-belief understanding [6]. In other words, theory-of-mind may be a necessary but not sufficient ability for success on this task to be achieved. It seems trivial to note that a 1-year-old child would fail the Sally–Anne task because of insufficient verbal ability. Several other hurdles have been suggested to stand in the way of passing the task that 3-year-olds might fall at, including the linguistic requirements [7], and the demands on executive resources, such as attention and inhibition [8,9]. Such processes could either mask expression of a representational theory of mind, or play a causal role in its development. The approaches that developmental psychologists have taken to address this issue can be broadly separated into two categories:

(i) Lowering task demands. Can the ability be demonstrated in a task that places fewer demands on the so-called 'performance factors'?
(ii) Explaining failure on the original task. Can the difference between individuals that pass and those that fail be positively linked to a difference in another process?

(i) *Lowering task demands*

A number of alternative tests of false-belief attribution have been devised that reduce some of the attentional and verbal requirements of the Sally–Anne task (see [6,10] for reviews). Tasks that require a physical rather than a verbal response, for example, by engaging children's readiness to help someone achieve their goal, have revealed evidence for false-belief attribution at 18 months [11]. Looking-time procedures remove response selection and language demands altogether, and have found positive results in children as young as 15 or even 12 months of age (e.g. infants look longer at a scene in which someone acts against the false belief that they should hold concerning an object's location or contents) [12]. Whether these results tap the same socio-cognitive skill that 4–5 year-olds use to solve the Sally–Anne task is a matter of debate; for example, the babies may expect people to search for objects where they saw them last and so react with surprise to the novelty of the event without explicit knowledge of belief [13], perhaps reflecting the existence of two systems for tracking belief [14]. Nevertheless, the possibility that younger children can reckon on the falsely held representations of other individuals suggests that some other process changes between the ages of 3 and 4 years to enable older children to succeed on the Sally–Anne task [10,15].

(ii) *Explaining failure on the original task*

Isolating the point of weakness with analogous tasks. Researchers have sought to investigate hypotheses attributing failure on the Sally–Anne task to mechanisms other than mental state representation by using tasks that are comparable but that lack the representational content. Birch & Bloom [16] suggest that what changes over the course of the third year is the ability to overcome the bias to report what is truly the case, or to overcome the 'curse of knowledge'. In order to respond correctly about where Sally will look for her marble, children have to repress their representation about where the marble actually is (the 'pull of the real'). Three-year old children fail other tasks that pose this difficulty but that do not contain representational content, such as describing 'false signs' after a scene has been changed [17]. Children at this age also have difficulties answering questions about counterfactual states of affairs, whether or not they contain references to beliefs (for example, if Sally had not moved the marble, where would it be now?), and interestingly the authors found a significant correlation between performance on counterfactual problems involving physical content and false-belief problems [18].

Correlations with other tasks. Another approach has capitalized on the fact that the performance on false-belief tasks is variable across individuals of the same age. Research has tested the same individuals both on theory-of-mind tasks and a suite of other tasks designed to tap a different process. In several studies, performance on false-belief tasks correlates with specific aspects of executive control, such as inhibiting a prepotent action plan in order to carry out a conflicting one (but not inhibiting over a delay) [19]. Multiple regression analyses reveal that this relationship exists over and above differences in verbal ability and age. Some authors reason that maturation in these executive processes plays a causal role in the development of a theory of mind, perhaps through allowing children to attend to the relevant features of social interactions in order to learn from them [9,19–22]. Others contend 3 year-olds possess knowledge of others as having beliefs and desires; the maturation of the executive processes is what allows older children to select the right response when required to think about beliefs that are in conflict with their own (reviewed in [15]). The point we want to make here is that, by showing that a significant amount of the variance in performance can be accounted for by variance in inhibitory skills, researchers have revealed an important role for this process in explaining the development of children's ability to reason about false beliefs. Similar analyses reveal a role for language development [23]. Differences between individuals at the same developmental period are a powerful tool for testing hypotheses about underlying mechanisms.

(b) *Object search tasks*

A similar focus on the cause of task failure has yielded interesting findings concerning children's knowledge

of the physical world of objects and their interactions. One finding that has struck researchers since Piaget is that pre-schoolers make some striking errors when searching for hidden objects that have undergone an invisible displacement, particularly if the final location must be inferred based on some physical principle, for example, that one object cannot pass through another. In one task conducted by Berthier *et al.* [24], children watch a ball roll down a ramp behind an occluder (see also [25]). A wall, clearly visible above the height of the occluder, will stop the ball's progress down the ramp and can be positioned at one of four locations, corresponding to four doors that children can open to search for the ball. Most children below the age of 3 do not use the position of the wall to infer where the ball will be at above chance levels. Do they lack the physical reasoning capacities that the test was designed to probe?

(i) *Lowering task demands*

The finding is particularly surprising because the task is based on a looking-time paradigm in which infants need only watch a display in which a ball is rolled towards a wall behind an occluder. The ball's final location is revealed in a location that is either consistent with physical principles (in front of a wall) or inconsistent with them (behind it). Children looked longer at the inconsistent display at four months of age [26]! A version of the ramp task in which a puppet did the searching also found longer looking to violation-of-expectation displays, when the puppet searched in the wrong place and found the ball, or searched in the right place and the ball was not there [27]. The implication is that children do have an expectation that one object cannot pass through another and are surprised when this principle is not adhered to, although some authors maintain that they may just be reacting to the perceptual novelty of the event [28]. However, other changes to the task aimed at lowering task demands but still requiring children to search—for example, increasing the salience of the wall through verbal cues or by using a human hand as the barrier—had no beneficial effect on children's performance [29].

(ii) *Explaining failure on the original task*

Isolating the point of weakness with analogous tasks. Mash *et al.* [30] removed the requirement for children to track the ball's trajectory and instead occluded the ball only after it had come to rest against the wall [30]. Revealingly, most 2- and 2.5-year-olds failed to locate the ball. However, in the 2.5-year-old age group, some individuals were successful; on trials in which children maintained unbroken eye-contact with the ball, search was correct on 90 per cent of trials. Focused visual attention was thereby implicated as an important factor that might also limit performance in the original task (see [31]). Many of the errors both in this task and the original resulted from perseverative reaching to the last location that had been searched; several individuals opened the same door on all 12 trials. It may be that, similar to the argument for false-belief tasks, the executive processes required for using information about the location of the ball to select a response (whether the source of that information is memory or inference), while inhibiting pre-potent responses based on pre-existing biases, prevents 2 year-olds from solving this task. Alternatively, 2 year-olds' knowledge of solidity and path-continuity may not be robust enough to support problem-solving. A correlationary study with different executive function tasks would be an interesting next step in addressing this question.

Tests designed to probe a particular cognitive ability naturally tax several mechanisms. This makes it difficult to interpret the difference between successful and unsuccessful individuals or groups. In this section, we have described two examples in which detailed analysis of what causes failure in psychological tasks has provided valuable insights into the cognitive mechanisms that undergo change during child development to enable successful social and physical problem-solving. The first approach that we have outlined is to pose the original conceptual task while attempting to minimize peripheral demands, to see if individuals that fail really lack the skill under study. This is important in the context of forming theories about evolutionary or ontogenetic change in this focal skill, where the lower limit is critical (note that because of the binary nature of the tasks we have discussed this analysis could be couched in terms of explaining 'success'; see [32] for discussion of the limitations of binary measures). The second approach in our classification could be seen as the inverse of the first: systematic variation of the peripheral task demands posed by the original task, in the absence of its specific conceptual content, to isolate the point of weakness. Here, individual differences can be valuable either when used within the context of analogous tasks (such as showing that individuals that fail the false-belief task are likely to fail the false-sign task); or alternatively when triangulating on mechanisms that influence the distribution of performance across individuals using task batteries designed to examine these skills from a different angle (in another context). Correlations between performance on one task and either an analogous one or on a task battery can only go so far in implicating a particular mechanism as the cause of failure. Directional relationships have been explored in developmental psychology using training studies and longitudinal analyses. Such an approach might also prove fruitful in comparative psychology, but is outside the scope of this article.

In §3, we aim to apply these approaches to comparative psychology, in particular, the question of the evolution of object concepts or 'folk physics'.

3. EXPLAINING CHIMPANZEES' FAILURE IN TOOL USE TASKS

Several non-human primate species use tools in their natural habitats and there has been a great deal of interest in the cognition underpinning their behaviour. If animals other than humans have an ability to represent objects and their abstract properties this might lend support to 'core cognition' accounts concerning the origin of concepts in humans. Recent research

has found that in a battery of tests apes (chimpanzees and orangutans) display similar physical knowledge to 2.5 year-old children, and the authors suggest that the important difference between humans and other apes lies in their social preferences and skills [33]. However, Penn *et al.* [34] contend that there are fundamental discontinuities between humans and other animals in their representational capacities, including those about the physical world. One line of evidence for this idea comes from a series of experiments conducted by Köhler [35] in which chimpanzees had to find new ways to gain out-of-reach food items. Köhler was struck by some of the errors that the chimpanzees made. Though they were sometimes surprisingly quick to solve novel problems, they often behaved as if they expected objects that were in perceptual contact to be connected [35]. For example, when stacking boxes in order to reach a banana attached to the ceiling, they were seen to lift the boxes and press them against the wall. This fits with the idea that non-human animals are limited to representations with 'first-order perceptual' content [34].

The results of a suite of studies run by Povinelli and colleagues were broadly consistent with the notion that chimpanzees use perceptually based information rather an abstract notion of object properties [36]. For example, in the trap tube task, in which the subject needs to push a piece of food out of a horizontal tube away from a trap, only one subject (out of five) learned to do so even when given over 100 trials in which to learn the solution. Why did the majority of individuals fail this task? Are these kinds of object relationships hard to grasp? Or could the task of avoiding the trap have been complicated by the requirement to use a tool to do so, which places additional demands on executive resources? Consistent with the latter explanation, Seed *et al.* [37] found that eight out of eight chimpanzees solved a version of the trap problem that did not require them to use a tool in under 100 trials [37]. This is an example of the first approach outlined earlier: lowering peripheral task demands. The performance of these subjects was then compared with naive subjects on a perceptually distinct transfer test made of new materials. Chimpanzees without experience on the first problem performed poorly on this task (only one subject was successful), but all of the three experienced subjects tested without a tool solved the new test in very few trials, suggesting that they had encoded information about the functional properties of the objects involved in the initial testing phase. Importantly, another group of experienced subjects tested with a tool required many more trials, and two out of four subjects were successful, revealing the critical importance of the manner of task presentation and the potentially confounding nature of the tool use variable. Further work is needed to isolate what it is about tool use that makes this task harder to solve; we suggest that using the second approach of correlation with analogous tasks could be valuable.

Povinelli and colleagues conducted another series of experiments that was directly aimed at testing Köhler's proposal that chimpanzees use perceptual contact rather than an abstract principle of mechanical connection when acting on objects. Subjects were required to use a stick tool to knock an out-of-reach apple down a slope and into reach. In each condition, there was both a 'connected' tool that was long enough to reach the apple, and a 'broken' tool composed of three small stick pieces that, when aligned, were the same length as the connected tool. In the staggered condition, the broken tool was arranged in front of the subjects so that although the pieces touched, they were staggered as shown in figure 1. In the aligned condition, the broken tool pieces were shown to the subject and then positioned on the ground in front of them so that they looked similar in length and form to the connected tool. The seven chimpanzees chose the correct (connected) tool significantly more often than the broken tool in the staggered condition, but in the aligned condition, performance was at chance in the four trials given. Povinelli and colleagues suggested two possible explanations. The first was that the chimpanzees' choices were indeed driven by the degree of perceptual contact rather than information about their mechanical properties. The second was that chimpanzees could not hold in mind which tool was broken when confronted with the illusion of contact, and so only passed the task when there was a visual reminder at the time of choice. To investigate further, the chimpanzees were given a choice between a connected but staggered tool and three pieces laid down end-to-end as before (figure 1).

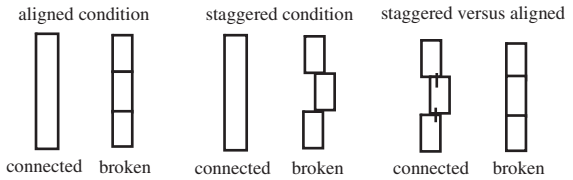

Figure 1. Touching stick experiment—Povinelli [36].

Chimpanzees preferred the aligned but unconnected tool, suggesting a reliance on the degree of perceptual contact over evidence concerning functional connection. Similarly, the apes failed to solve other tasks in which there was no perceptual information about which option was unconnected at the time of choice; for example, they failed to choose a complete rake to bring food into reach over one that was laid down in two unconnected pieces and therefore non-functional, even over the course of 12 trials (although one chimpanzee's performance did approach significance). The authors interpreted the results as 'strongly confirming Köhler's idea that the optics of the situation tends to control the apes' behaviour' (Povinelli [36, p. 252]). However, we think that the alternative explanation outlined by the authors could still explain the results. The follow-up study reveals that the degree of perceptual contact certainly influences chimpanzees' choices. However, when two options are equally good or bad, adult humans also show significant biases towards options with a greater degree of perceptual contact [38]. If the chimpanzees lacked sufficient focused attention or working-memory resources to mark or to hold in mind which of the tools was broken, then their choices might well reflect a similar bias. Therefore, the result of the follow-up experiment should not be taken as concrete

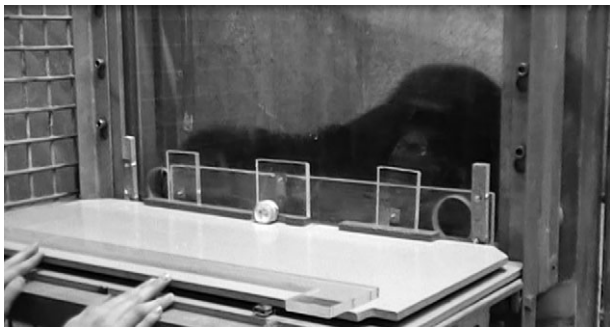

Figure 2. Experimental set-up for the aligned tool study showing the visible condition. The subject has selected the connected tool. Note that the hole in front of the other choice is now closed off by the sliding panel.

evidence for the 'Köhlerian view'—that chimpanzees do not have a notion of connection any deeper than mere contact.

4. EXPERIMENTAL STUDY: THE ALIGNED TOOL TASK REVISITED

In the following study, we set out to investigate the cause of failure on the aligned tool problem, by taking two of the approaches outlined in the earlier-mentioned review: lowering peripheral task demands, and comparing with the results of an analogous task. The experimenters in the original experiment spent a total of 30 s manipulating the two tools before setting them down and allowing the chimpanzees to make their choice, giving the chimpanzees ample information about which tool was broken, but also imposing quite a long delay over which they might become distracted. We aimed to reduce the task demands that might obscure a chimpanzee's ability to solve the task. Attending to the distant actions of a human experimenter might be problematic, because in the lives of captive chimpanzees humans carry out many actions that are irrelevant to a chimpanzee's chance to obtain food. Rather than demonstrating the properties of the tools, we just laid them in position piece by piece, doing away with the 30 s period of tool demonstration which might have led to the chimpanzees becoming frustrated. The tools were placed on a table parallel to the window such that each ran towards a single reward, hooked onto both of the tools, in the middle of the table (figure 2). Subjects could make a choice by moving a sliding panel to open one of two holes in the window (in front of the ends of the tools). Only by taking the complete instead of the broken tool could subjects bring the reward in an arc across the table, through the hole and into reach. There was no training phase with the complete tool; instead, the chimpanzees were confronted with a choice between the two options on their first trial.

In the *visible* condition, the 6 cm gap between the pieces of the broken tool could be seen at all times. In the *aligned* condition, broken tool was set down in two pieces as in the visible condition, and then ends were pushed together before the chimpanzees were allowed to choose. We also included a condition that was analogous to the aligned condition in terms of the demands placed on attention and memory, but without the appearance of perceptual contact that might 'tend to control' the apes' behaviour. In this *covered* condition, the middle section of both tools was covered over with a small occluder before the chimpanzees could make their choice. The latter two conditions therefore presented the same demands on executive processes, namely to attend to the tool set-up, and remember which tool was broken over a brief delay, but only the aligned condition featured perceptual contact between the broken pieces at the time of choice.

Table 1. Name, age, sex, rearing history and experiments in which each subject participated.

name	age (years)	sex	rearing history	experiment participation
Robert	35	male	nursery	1, 2
Corrie	34	female	nursery	1, 2
Fraukje	34	female	nursery	1, 2
Riet	33	female	nursery	1, 2
Dorien	30	female	nursery	1, 2
Natascha	30	female	nursery	1, 2
Sandra	17	female	mother	1, 3
Frodo	17	male	mother	1, 3
Swela	15	female	mother	1
Patrick	13	male	mother	1, 3
Pia	11	female	mother	1, 3
Lome	9	male	mother	1, 2, 3
Tai	8	female	mother	1, 2, 3
Lobo	6	male	mother	1, 2
Kofi	5	male	mother	1, 2
Kara	5	female	mother	1, 2, 3

(a) *Method*

(i) *Subjects*

Chimpanzees (*Pan troglodytes;* $n = 16$) housed at the Wolfgang Köhler Research Centre, Leipzig Zoo (Leipzig, Germany) participated in this experiment. All subjects lived as a social group, with access to indoor and outdoor areas. Subjects were tested individually in a familiar indoor testing room. Water was freely available and subjects were not food-deprived for testing. Table 1 shows the age, sex, rearing history and experimental participation of each subject.

(ii) *Materials*

The experiments were carried out on a grey plastic table (80 × 39 cm) with a sliding plastic table on top of it (78 × 35 cm). The table was attached to a metal L-frame in front of a Plexiglas testing window (69 × 48 cm) with two holes at hand height (6 cm in diameter). A sliding Plexiglas panel was attached to the testing window and had to be moved to the right or left by the subject before they could reach through the hand holes. Once moved to one side, the sliding panel blocked the other hole, restricting subjects to one choice. In each condition, one option was a 'connected' tool (24 × 1.5 × 1.5 cm), and the other was a 'broken' one made of two pieces; one hook (6 cm long) and one end piece (visible and covered—12 cm long, aligned—18 cm long). In the covered condition, two

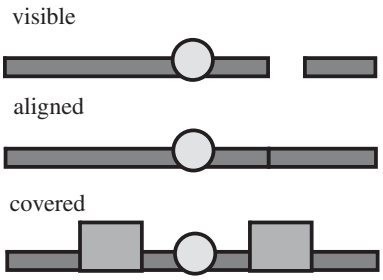

Figure 3. Experiment 1—conditions. Appearance of the tools at the time of choice for each condition (visible, aligned and covered) is shown. The correct, connected tool is shown on the left.

green plastic covers were used (made of two square pieces of plastic 10 × 10 cm taped together to form a 'billboard'). In addition, one large green plastic occluder (80 × 30 cm) was used in all conditions.

(iii) *Procedure and design*

The experimenter sat facing the subject behind the table. Each session began with a pre-test procedure of placing four grapes, one at a time in front of the holes in the Plexiglas window (two on each side). This was to ensure that subjects knew how to use the sliding panel, and that picking from either side could result in reward. The experimenter set up the tools from left to right in parallel with the window on the sliding table out of reach of the subject, calling the subject's name to attract attention if necessary. The banana slice was then hooked in the middle onto both of the tools. In the visible condition, the tool pieces were placed on the sliding table so that there was a 6 cm gap between the handle and the hook of the broken tool (figure 2).

In the aligned condition, the tool pieces were positioned and the reward placed in the same way as the visible condition, and then the end of the broken tool was pushed so that the gap was closed and no longer visible. The resulting appearance was of a tool the same shape and size as the connected tool. The experimenter also placed her hand on the end of the connected tool to avoid side bias owing to local enhancement (tools were always manipulated from left to right). In the covered condition, after the reward was attached, the centre of the two tools was covered by the 10 cm 'billboard' cover, such that the gap in the broken tool could no longer be seen, and both tools looked visually identical at time of choice. Figure 3 shows the three conditions.

In each condition, the procedure took 15 s, after which the experimenter placed an occluder over the table for a period of approximately 3 s. The occluder was then lifted and the table slid towards the subject. The sliding Plexiglas panel was then released and could be moved to one side. Subjects were only able to obtain the food reward if they chose the tool connected to the banana; if the 'broken' tool was chosen the table was slid away from the subject and the food reward was removed by the experimenter. Each testing session consisted of 12 trials, four of each condition. There were three sessions in total for each subject, making a total of 36 trials, 12 of each condition. The order of the conditions was randomized between trials with the constraint that the same condition was not presented more than twice in a row. The location of the broken tool (left versus right) was counterbalanced so that it appeared on each side for each condition an equal number of times, but not on the same side more than twice in a row.

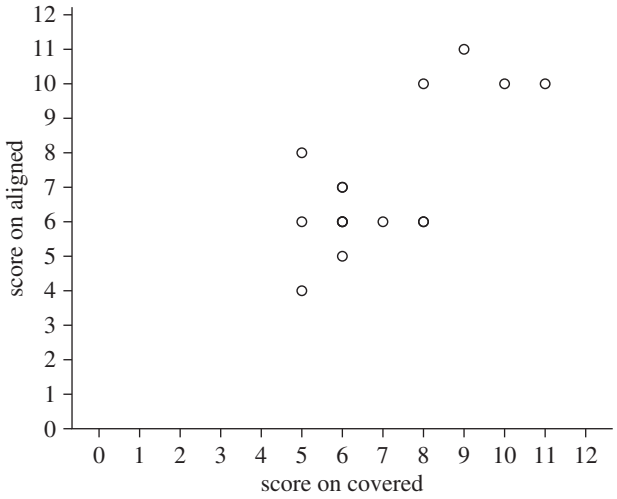

Figure 4. Scatterplot showing performance on the aligned and covered conditions of experiment 1.

(iv) *Scoring and analysis*

Subjects' choices were recorded live. A correct choice was recorded if the subject chose the connected tool. All trials were video-recorded and a second observer scored 100 per cent of the trials. Inter-observer reliability was 100 per cent. Data were analysed using IBM SPSS v. 19, using two-tailed tests with alpha set at 0.05. We used the exact Wilcoxon sign-ranks tests when the sample size was small ($n < 15$) [39].

(b) *Results and discussion*

Chimpanzees performed best in the visible condition: a Friedman test revealed significant differences between conditions ($\chi^2_{(\text{d.f.}=2,\ n=16)} = 8.59$, $p = 0.014$) and pairwise comparisons revealed that subjects performed significantly better in the visible condition than the aligned condition (Wilcoxon test, $T = 3$, $n = 12$, $p < 0.02$) and the covered condition ($T = 7.5$, $n = 13$, $p < 0.02$). There was no significant difference between the aligned and the covered condition ($T = 35$, $n = 12$, n.s.).

Six individuals were above chance according to a binomial test (scored 10/12 or more correct) on the visible condition. Four of these six also passed the aligned condition, and two of those four passed all three conditions (the other two scored eight and nine out of 12). None of the chimpanzees passed only the covered and/or the aligned tasks. A partial correlation controlling for age, and performance on the visible condition, revealed a significant relationship between performance on the covered and the aligned conditions ($n = 16$, $r = 0.605$, $p = 0.022$; figure 4).

The experiment revealed that in this larger sample of 16 individuals, four chimpanzees were capable of

Table 2. Means, standard deviations and test statistics on each condition for each experiment. Number of successful individuals on each condition, according to a binomial test, is shown in brackets after the mean. Wilcoxon test statistic and one-tailed significance level are also shown.

experiment	*n*	condition	mean	s.d.	*T*	*p*-value
1	16	visible	8.81 (10)	1.94	1.00	0.001
		aligned	7.13 (4)	2.06	6.50	0.05
		covered	7.00 (3)	1.83	7.5	0.025
2	11	visible	9.45	2.25	1.50	0.01
		aligned	7.18	1.89	7.50	0.05
		aligned no occlusion	7.91	1.51	1.50	0.01
		aligned no occlusion, no delay	7.18	2.36	18.00	n.s.
3	7	aligned	9.86	1.68	0	0.01
		negative	6.86	1.57	2.50	n.s.
		inductive	6.86	2.04	5.50	n.s.
		positive	9.29	0.95	0	0.01

solving a broken tool task in which both tools were in perceptual contact with the reward at the time of choice. However, as in the original experiment run by Povinelli [36], as a group, chimpanzees performed better on the visible condition. Revealingly, there was a correlation between performance on the aligned task and an analogous task that did not require chimpanzees to overcome the perceptual impression of connection, but which did require them to pay focused attention to an experimenter's demonstration, and to remember what they has seen for a short delay. This suggested that these executive processes could play a role in determining success or failure on the aligned task. This relationship existed even when age and performance on the visible task (which could be taken as a metric for motivation to succeed) was controlled for.

5. FOLLOW-UP STUDIES

(a) *Unsuccessful subjects: further reductions to task requirements*

We aimed to investigate the role of memory by reducing the delay over which chimpanzees had to hold in mind which tool was broken. Working-memory only has limited temporal capacity, so reducing the delay could improve performance [40]. In humans, eye movements have been shown to be a mechanism for rehearsing visuo-spatial working memory [41]; therefore, occluding the tools after they are aligned could distract attention and prevent eye movements necessary for rehearsal of the broken tool location. We therefore removed the occlusion phase in order to further reduce task requirements. We predicted that reducing delay and removing occlusion could improve performance compared with the aligned condition from experiment 1.

(i) *Methods*

We gave 11 of the 12 subjects that failed the aligned condition (one declined to participate further) 12 trials of the original visible and aligned conditions, and 12 trials of each of two modified versions of the aligned condition in which we removed the occlusion phase of the set-up. In one (*aligned—no occlusion*), we kept the total delay between tool set-up and choice constant with the original aligned condition; in the other, we removed the delay and immediately pushed the table towards the subjects after aligning the broken tool (*aligned—no occlusion no delay*). Subjects received the four conditions in an interleaved fashion with counterbalancing as in the previous experiment, with 12 trials per day for a total of 48 trials.

(ii) *Results and discussion*

Performance remained higher in the visible condition than any of the aligned conditions (table 2). Pairwise comparisons revealed that subjects performed significantly better in the visible condition than in the aligned (Wilcoxon test, $T = 2.5$, $n = 10$, $p < 0.02$) and aligned—*no occlusion, no delay* condition ($T = 0$, $n = 10$, $p < 0.002$). There was no significant difference with the aligned—*no occlusion* condition ($T = 9.5$, $n = 10$, $p > 0.05$). We found no statistically significant differences between performances on the three aligned conditions (all pairwise comparisons n.s.). The study therefore did not find support for the idea that failure on the aligned condition was caused by working-memory deficits that could be alleviated by reducing the delay or removing the occlusion phase. However, working-memory overload cannot be ruled out as a contributor to failure on the aligned and covered tasks, as it could result from the number of mental representations required at a single point in time. In the aligned condition, the subject has to hold in mind the representation that one of the tools, though seemingly connected to the reward, is not, while choosing between two options that appear identical. In the visible condition and Povinelli's [36] staggered condition, there is a constant visual reminder consistent with reality that one of the tools was broken. Further work could look at correlations with other tasks designed to tap working-memory and other executive functions such as attentional focus to look for positive evidence of an association between performance on this task and specific executive functions.

It was notable that three subjects were significantly above chance on the aligned condition in this study. They were included along with the successful group from experiment 1 in the next follow-up experiment.

(b) *Successful subjects: does the gap need to be seen to be believed?*

The aim of the experiment described already was to investigate the cause of failure on the aligned tool task. However, we found that four subjects passed the aligned condition in the initial 12 trials, and three more achieved significance in the follow-up condition. In the second follow-up, we aimed to look more closely at what these results mean. Should we reject the Köhlerian view and accept that these subjects can represent connection as a mechanical rather than a perceptual property? Perhaps an alternative explanation for their success stems from a corollary of our explanation for why some individuals failed. If individuals paid close attention to the hiding events and remembered where they had seen a gap, they could have avoided this tool without knowing the functional relevance of the gap. In this follow-up experiment, we removed this perceptual cue to a correct choice, and required subjects to infer which tool was connected from evidence about each tool's ability to function in moving the reward.

(i) *Methods*

Subjects were tested on the aligned condition and three inference conditions. In all of the inference conditions, the experimenter set up the tools underneath the occluder so that the subject could not see the initial tool placement. The tools were set up as in the covered condition of experiment 1, with the 10 cm covers hiding the gap in the broken tool and the corresponding section of the connected tool. The occluder was then removed and the experimenter manipulated the tools from left to right, but the covers made it impossible for the subject to see which tool was broken and which was connected directly. Then the sliding table was pushed forward to allow the subjects to make their choice. The tools were manipulated so as to provide three different types of information from which the subjects could infer which tool was connected.

(ii) *Positive: if the tool moves the reward then it is connected to it*

The experimenter moved the end of each tool four times from left to right. The connected tool moved the banana, but the broken tool did not.

(iii) *Negative: if the tool does not move the reward then it is not connected to it*

The experimenter moved the end of the broken tool outwards by approximately 4 cm. She also touched the end of connected tool it to avoid side bias. The broken tool moved, but did not move the banana, the connected tool did not move.

(iv) *Inductive: which tool is most likely to be connected?*

The experimenter moved the outer end of the broken tool back-and-forth with one hand, and the banana and the hook end with the other hand at the same time. To move the banana, she reached around the connected tool, hiding its movement from the view of the subject with her arm. She moved the end of the connected tool back and forth, which moved the banana as in the *positive* condition. The subject should infer that the tool which moved the banana without the experimenter's hand on it is the most probable to be connected to it.

Each testing session consisted of 12 trials, with a total of four testing sessions (48 trials), 12 trials for each condition. The order of the conditions and location of the broken tool (left versus right) were randomized and counterbalanced as in experiment 1.

(iv) *Results and discussion*

As a group, the seven chimpanzees were significantly above chance in the aligned and positive conditions, but not the *negative* and *inductive* conditions (see table 2 for means and test statistics).

Four chimpanzees were able to use indirect evidence to infer the location of the correct tool in the positive condition (scored 10/12 or more). One chimpanzee solved both the negative and inductive conditions, but puzzlingly only scored 8/12 on the aligned and positive condition.

The performance of the chimpanzees on the positive condition reveals that they did not need to see the gap in the broken tool (or indeed a perceptually continuous tool) in order discriminate between the options. This supports the view that chimpanzees know more about connection than can be gleaned from perceptual forms alone. However, they failed two inference conditions that could be said to be more complex, reasoning by exclusion (basing the decision on what is not the case), and induction (inferring which of two options is more likely when both are possible). The results are consistent with previous work on inference, which has shown that apes find these types of inference difficult. Call [42] presented apes with a task which required them to choose which of two cups was baited with food based on exclusion. In the auditory condition, where apes were shown that one cup made no sound when it was shaken, only one bonobo and two gorillas were successful, though two chimpanzees and two orangutans passed this task in another study [43]. In another task, apes were able to infer which of two boards had a food reward beneath it when one was inclined and one lay flat on the table (analogous to our positive condition) [44]. Control tasks showed that their behaviour was not driven by a simple preference for slanted boards. However, when both boards were inclined, but only one of them was visibly supported by a wooden strut, none of the apes were able to select the unsupported board as being more likely to have food beneath it. Our inductive condition has the same logical structure: both options are possible but only one *must* be connected. One interpretation of our results is therefore that apes are capable of making inferences based on the logic that follows from causal relationships between objects, but not inferences requiring more abstract logical operations. This question is worthy of further study. However, an alternative explanation of our results is that they solved the only condition in which there was an asymmetry in the movement of the food reward because of a pre-existing bias (unlike the positive condition, in the negative inference condition neither tool was associated with food movement,

in the inductive condition both tools were). A third possibility is that the movement of the tools in the negative and aligned conditions was less attention-grabbing and more difficult to follow for the apes, thus placing higher demands on executive resources, which this study has already implicated as a limiting factor on performance. Future work will be necessary to disambiguate these alternatives.

This paper has focused on the challenge of interpreting negative results in the cognitive sciences. We have identified a number of approaches that have been used to face this challenge. By lowering task demands, experimenters can find evidence for the ability that the original task was designed to diagnose. By conducting analogous tasks that pose the same demands on other processes but which remove the demand that was originally in focus, the role of other processes and biases can be uncovered. Perhaps most promisingly, individual variation in performance can be linked to individual variation either in analogous tasks, or in batteries designed to isolate different processes, to provide positive evidence for an influence of the latter on the former. We have seen that these approaches have provided support for the role of specific executive functions in the development of social and physical problem-solving in young children. We also report evidence that executive functions are a limiting factor that may determine whether chimpanzees pass or fail tool use tasks. These results suggest that failure on some of the most widely used tasks may not provide evidence of a lack in the ability that they were designed to test. Instead, the ability may be present but masked by differences in executive function. Alternatively, the difference between individuals that pass and fail may be better characterized by differences in a broader, domain general mechanism. It may be time for the 'peripheral' processes that are known to influence performance to be brought back into focus. Differences in these mechanisms may have played an important role in primate evolution, and may explain changes in behaviour and problem-solving ability over the course of development. Combining the inter-individual variation approach with an inter-specific comparative approach can therefore reveal important clues about cognitive evolution (see also [45]).

REFERENCES

1 Heyes, C. 2012 Simple minds: a qualified defence of associative learning. *Phil. Trans. R. Soc. B* **367**, 2695–2703. (doi:10.1098/rstb.2012.0217)

2 Wimmer, H. & Perner, J. 1983 Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition* **13**, 103–128. (doi:10.1016/0010-0277(83)90004-5)

3 Baron-Cohen, S., Leslie, A. M. & Frith, U. 1985 Does the autistic child have a 'theory of mind'? *Cognition* **21**, 37–46. (doi:10.1016/0010-0277(85)90022-8)

4 Wellman, H. M., Cross, D. & Watson, J. 2001 Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* **72**, 655–684. (doi:10.1111/1467-8624.00304)

5 Penn, D. C. & Povinelli, D. J. 2007 On the lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Phil. Trans. R. Soc. B* **362**, 731–744. (doi:10.1098/rstb.2006.2023)

6 Bloom, P. & German, T. P. 2000 Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* **77**, B25–B31. (doi:10.1016/S0010-0277(00)00096-2)

7 de Villiers, J. G. & de Villiers, P. A. 2000 Linguistic determinism and the understanding of false beliefs. In *Children's reasoning and the mind*, pp. 191–228. Hove, UK: Psychology Press.

8 Hala, S. & Russell, J. 2001 Executive control within strategic deception: a window on early cognitive development? *J. Exp. Child Psychol.* **80**, 112–141. (doi:10.1006/jecp.2000.2627)

9 Perner, J., Lang, B. & Kloo, D. 2002 Theory of mind and self-control: more than a common problem of inhibition. *Child Dev.* **73**, 752–767. (doi:10.1111/1467-8624.00436)

10 Baillargeon, R., Scott, R. M. & He, Z. 2010 False-belief understanding in infants. *Trends Cogn. Sci.* **14**, 110–118. (doi:10.1016/j.tics.2009.12.006)

11 Buttelmann, D., Carpenter, M. & Tomasello, M. 2009 Eighteen-month-old infants show false belief understanding in an active helping paradigm. *Cognition* **112**, 337–342. (doi:10.1016/j.cognition.2009.05.006)

12 Onishi, K. H. & Baillargeon, R. 2005 Do 15-month-old infants understand false beliefs? *Science* **308**, 255–258. (doi:10.1126/science.1106480)

13 Perner, J. & Ruffman, T. 2005 Infants' insight into the mind: how deep? *Science* **308**, 214–216. (doi:10.1126/science.1111656)

14 Apperly, I. A. & Butterfill, S. A. 2009 Do humans have two systems to track beliefs and belief-like states? *Psychol. Rev.* **116**, 953–970. (doi:10.1037/a0016923)

15 Leslie, A. M., Friedman, O. & German, T. P. 2004 Core mechanisms in 'theory of mind'. *Trends Cogn. Sci.* **8**, 528–533. (doi:10.1016/j.tics.2004.10.001)

16 Birch, S. A. J. & Bloom, P. 2004 Understanding children's and adult's limitations in mental state reasoning. *Trends Cogn. Sci.* **8**, 255–260. (doi:10.1016/j.tics.2004.04.011)

17 Leekam, S., Perner, J., Healey, L. & Sewell, C. 2008 False signs and the non-specificity of theory of mind: evidence that preschoolers have general difficulties in understanding representations. *Br. J. Dev. Psychol.* **26**, 485–497. (doi:10.1348/026151007×260154)

18 Riggs, K. J., Peterson, D. M., Robinson, E. J. & Mitchell, P. 1998 Are errors in false belief tasks symptomatic of a broader difficulty with counterfactuality? *Cogn. Dev.* **13**, 73–90. (doi:10.1016/s0885-2014(98)90021-1)

19 Carlson, S. M., Moses, L. J. & Claxton, L. J. 2004 Individual differences in executive functioning and theory of mind: an investigation of inhibitory control and planning ability. *J. Exp. Child Psychol.* **87**, 299–319. (doi:10.1016/j.jecp.2004.01.002)

20 Carlson, S. M. & Moses, L. J. 2001 Individual differences in inhibitory control and children's theory of mind. *Child Dev.* **72**, 1032–1053. (doi:10.1111/1467-8624.00333)

21 Russell, J., Mauthner, N., Sharpe, S. & Tidswell, T. 1991 The 'Windows Task; as a measure of strategic deception in preschoolers and autistic subjects. *Br. J. Dev. Psychol.* **9**, 331–349. (doi:10.1111/j.2044-835X.1991.tb00881.x)

22 Hughes, C. & Ensor, R. 2007 Executive function and theory of mind: predictive relations from ages 2 to 4. *Dev. Psychol.* **43**, 1447–1459. (doi:10.1037/0012-1649.43.6.1447)

23 Milligan, K., Astington, J. W. & Dack, L. A. 2007 Language and theory of mind: meta-analysis of the relation between language ability and false-belief understanding. *Child Dev.* **78**, 622–646. (doi:10.1111/j.1467-8624.2007.01018.x)

24 Berthier, N. E., DeBlois, S., Poirer, C. R., Novak, M. A. & Clifton, R. K. 2000 Where's the ball? Two- and three-year-olds reason about unseen events. *Dev. Psychol.* **36**, 394–401. (doi:10.1037/0012-1649.36.3.394)

25 Hood, B., Carey, S. & Prasada, S. 2000 Predicting the outcomes of physical events: two-year-olds fail to reveal knowledge of solidity and support. *Child Dev.* **71**, 1540–1554. (doi:10.1111/1467-8624.00247)

26 Spelke, E. S., Breinlinger, K., Macomber, J. & Jacobson, K. 1992 Origins of knowledge. *Psychol. Rev.* **99**, 605–632. (doi:10.1037/0033-295X.99.4.605)

27 Mash, C., Novak, E., Berthier, N. E. & Keen, R. 2006 What do two-year-olds understand about hidden-object events? *Dev. Psychol.* **42**, 263–271. (doi:10.1037/0012-1649.42.2.263)

28 Hood, B. M. 2004 Is looking good enough or does it beggar belief? *Dev. Sci.* **7**, 415–417. (doi:10.1111/j.1467-7687.2004.00358.x)

29 Keen, R., Berthier, N., Sylvia, M. R., Butler, S., Prunty, P. K. & Baker, R. K. 2008 Toddlers' use of cues in a search task. *Infant Child Dev.* **17**, 249–267. (doi:10.1002/icd.550)

30 Mash, C., Keen, R. & Berthier, N. E. 2003 Visual access and attention in two-year-olds' event reasoning and object search. *Infancy* **4**, 371–388. (doi:10.1207/s15327078in0403_04)

31 Butler, S. C., Berthier, N. E. & Clifton, R. K. 2002 Two-year-olds' search strategies and visual tracking in a hidden displacement task. *Dev. Psychol.* **38**, 581–590. (doi:10.1037/0012-1649.38.4.581)

32 Thornton, A. & Lukas, D. 2012 Individual variation in cognitive performance: developmental and evolutionary perspectives. *Phil. Trans. R. Soc. B* **367**, 2773–2783. (doi:10.1098/rstb.2012.0214)

33 Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B. & Tomasello, M. 2007 Humans have evolved specialized skills of social cognition: the cultural intelligence hypothesis. *Science* **317**, 1360–1366. (doi:10.1126/science.1146282)

34 Penn, D. C., Holyoak, K. J. & Povinelli, D. J. 2008 Darwin's mistake: explaining the discontinuity between human and nonhuman minds. *Behav. Brain Sci.* **31**, 109–130. (doi:10.1017/S0140525X08003543)

35 Köhler, W. 1926 *The mentality of apes*. 2nd edn. New York, NY: Harcourt, Brace & Company.

36 Povinelli, D. J. 2000 *Folk physics for apes: the chimpanzee's theory of how the world works*. Oxford, UK: Oxford University Press.

37 Seed, A. M., Call, J., Emery, N. J. & Clayton, N. S. 2009 Chimpanzees solve the trap problem when the confound of tool use is removed. *J. Exp. Psychol. Anim. Behav. Process.* **35**, 23–34. (doi:10.1037/a0012925)

38 Silva, F. J., Silva, K. M., Cover, K. R., Leslie, A. M. & Rubalcaba, M. A. 2008 Humans' folk physics is sensitive to physical connection and contact between a tool and reward. *Behav. Process.* **77**, 327–333. (doi:10.1016/j.beproc.2007.08.001)

39 Mundry, R. & Fischer, J. 1998 Use of statistical programs for nonparametric tests of small samples often leads to incorrect *p* values: examples from animal behaviour. *Anim. Behav.* **56**, 256–259. (doi:10.1006/anbe.1998.0756)

40 Baddeley, A. D. 2010 Working memory. *Curr. Biol.* **20**, R136–R140. (doi:10.1016/j.cub.2009.12.014)

41 Lawrence, B., Myerson, J. & Abrams, R. 2004 Interference with spatial working memory: an eye movement is more than a shift of attention. *Psychon. Bull. Rev.* **11**, 488–494. (doi:10.3758/bf03196600)

42 Call, J. 2004 Inferences about the location of food in the great apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, and *Pongo pygmaeus*). *J. Comp. Psychol.* **118**, 232–241. (doi:10.1037/0735-7036.118.2.232)

43 Hill, A., Collier-Baker, E. & Suddendorf, T. 2011 Inferential reasoning by exclusion in great apes, lesser apes, and spider monkeys. *J. Comp. Psychol.* **125**, 91–103. (doi:10.1037/a0020867)

44 Call, J. 2007 Apes know that hidden objects can affect the orientation of other objects. *Cognition* **105**, 1–25. (doi:10.1016/j.cognition.2006.08.004)

45 Herrmann, E. & Call, J. 2012 Are there geniuses among the apes? *Phil. Trans. R. Soc. B* **367**, 2753–2761. (doi:10.1098/rstb.2012.0191)