Developmental Science 9:3 (2006), pp 323–333

PAPER

Choose and choose again: appearance–reality errors, pragmatics and logical ability

Gedeon O. Deák and Brian Enright

Department of Cognitive Science, University of California, San Diego, USA

Abstract

In the Appearance/Reality (AR) task some 3- and 4-year-old children make perseverative errors: they choose the same word for the appearance and the function of a deceptive object. Are these errors specific to the AR task, or signs of a general question-answering problem? Preschoolers completed five tasks: AR; simple successive forced-choice question pairs (QP); flexible naming of objects (FN); working memory (WM) span; and indeterminacy detection (ID). AR errors correlated with QP errors. Insensitivity to indeterminacy predicted perseveration in both tasks. Neither WM span nor flexible naming predicted other measures. Age predicted sensitivity to indeterminacy. These findings suggest that AR tests measure a pragmatic understanding; specifically, different questions about a topic usually call for different answers. This understanding is related to the ability to detect indeterminacy of each question in a series. AR errors are unrelated to the ability to represent an object as belonging to multiple categories, to working memory span, or to inhibiting previously activated words.

Introduction

In a widely replicated *Appearance/Reality* (AR) test (Flavell, Flavell & Green, 1983), children see a deceptive object (e.g. a magnet that looks like candy), learn its real function and then answer two questions. One question is about its appearance ('What does it look like; a magnet or chocolate?'), the other is about its real function ('What is it really, a magnet or chocolate?'). Many 3- and 4-year-olds respond by answering 'magnet' to both questions, or 'candy' to both questions. This type of error often has been assumed to reflect a conceptual deficit, for example inability to represent two categories for an object (Flavell *et al.*, 1983), or inability to represent the child's own changing beliefs about its identity (Gopnik & Astington, 1988).

From another perspective, AR errors are cases of *perseveration*, or inappropriate repetition of an earlier response. In the AR task children choose a word to correctly answer one question, but then sometimes choose it again to answer a different question. This might stem from a general tendency to perseverate. Two- to 4-year-olds perseverate in other tests that pose multiple questions about an object. For example, when asked to sort cards by one rule (e.g. 'All blue things go here . . . all red things go there'), then switch and sort the same cards by a different rule ('all cars go here . . . all flowers go there'), most 3-year-olds and some 4-year-olds continue to use the first rule (Frye, Zelazo & Palfai, 1996). In another test (Deák, 2000) 3- and 4-year-olds hear several words for an object, with different phrases implying different meanings for each word. Some children nevertheless infer that each word refers to the same property (e.g. the object's substance). In both tasks, then, many 3-year-olds and some 4-year-olds inappropriately repeat responses, as in the AR test.

This similarity was explored by Deák, Ray and Brenneman (2003). They found that children who make AR errors also perseverate when answering two forced-choice questions about non-deceptive objects. This suggests that AR errors come from a general tendency to perseverate in answering questions, not a specific problem with representing deceptive objects, or with grasping the conceptual implications of the AR questions. Note that if there is a general perseverative tendency it seems to be limited to verbal choices (i.e. forced choice between words or phrases). Sapp, Lee and Muir (2000) replicated the usual perseverative errors in a standard verbal AR test, but not in a non-verbal test where children made choices between objects (see also Rice, Koinis, Sullivan, Tager-Flusberg & Winner, 1997).

The current study tests several alternative hypotheses about why children make AR errors. Three- and 4-year-

Address for correspondence: Gedeon O. Deák, Department of Cognitive Science, University of California, San Diego, 9500 Gilman Dr., La Jolla, California 92093-0515, USA; e-mail: deak@cogsci.ucsd.edu

old children completed a standard object-identity AR test, plus other relevant tests. The AR test is sometimes used as a measure of theory-of-mind development (e.g. Andrews, Halford, Bunch, Bowden & Jones, 2003; Carlson, Moses & Breton, 2002), but this usage is problematic (e.g. Deák *et al.*, 2003; Hansen & Markman, 2005). In this study we focus on several other hypotheses about AR errors. These hypotheses relate AR errors to children's conceptual flexibility, their executive cognitive processing capacities and their pragmatic knowledge and logical awareness. The hypotheses are:

1. Representational inflexibility: Young children might have trouble keeping in mind more than one representation of a thing. That is, when they associate a category or label with a thing, they cannot easily activate a second category or label (Flavell *et al.*, 1983; Flavell, Green & Flavell, 1986; Melot & Houdé, 1998). For example, it has been reported that children claim a character cannot be both 'father' and 'doctor' (Sigel, Saltz & Roskind, 1967). This finding is, however, spurious: 2- to 4-year-olds readily accept several labels for a story character (Clark & Svaib, 1997; Deák & Maratsos, 1998). More pertinent to AR studies, 3-year-olds readily produce several words for objects with distinct appearance and function (e.g. 'dog' and 'puppet' for a dog puppet: Deák & Maratsos, 1998; Deák, Yen & Pettit, 2001). This capacity for 'flexible naming' is robust: for example, children maintain that all the words they produced for an object are correct (Deák & Maratsos, 1998), ruling out the possibility that they merely forget the last label they produced, or capriciously change their minds about which label is correct. Either of these possibilities would indicate representational fickleness but not flexibility.

Though flexible naming by 3-year-olds seems to contradict representational inflexibility, it has not been shown that the same children who make AR errors also can flexibly label complex objects. Without showing this we cannot rule out sampling differences between studies (e.g. children in flexible naming studies have had superior cognitive skills). However, if the same children who make AR errors also show flexible naming of similar objects, it would disconfirm representational inflexibility accounts of AR errors. To test this idea, children in the current study completed the *Flexible Naming* test from Deák and Maratsos (1998). In this test children saw *representational objects* like a dog puppet, and conversational prompts were used to elicit several words per object, including appearance and function labels. Children were also asked to verify that each pair of labels was simultaneously appropriate, to rule out forgetting or changing their minds. Finally, children were asked whether each word refers to the object's appearance or function (Deák *et al.*, 2001). These follow-up questions serve as an alternative appearance–reality test, and address other hypotheses outlined below.

2. Working memory (WM) capacity: An alternative to representational inflexibility is the hypothesis that immaturity of some executive cognitive capacities cause AR errors. One candidate is working memory capacity (Rice *et al.*, 1997). Preschoolers' verbal WM span is small, and it increases between 3 and 5 years (Gathercole & Adams, 1993). The AR task requires children to keep a question and two words in memory. When they are choosing between the words, the question (i.e. 'looks like . . .' or 'is really . . .') might not remain in WM. Perseveration would then be a reasonable fall-back response: if you cannot recall the question, repeat the last answer, which is still primed and accessible (see Gershkoff-Stowe, 2002). Support for this idea is mixed: Bialystok and Senman (2004) found a correlation of $r = .35$ ($R^2 = .12$) between digit span (a WM capacity measure) and AR errors. However, they did not control for age or verbal skills, so the finding is ambiguous. By contrast, Carlson *et al.* (2002) found a mean partial $r = .03$ between AR and verbal span when age and IQ were partialled out, indicating no specific relation. Also, Deák *et al.* (2003) found no correlation between AR and word span when age and vocabulary were controlled. However, a floor effect in the word span test might have reduced the association. Also, digit and word span tests used in prior studies depend partly on vocabulary, which also correlates with AR test performance (Deák *et al.*, 2003). Non-word span is an alternative measure of WM capacity that is less related to vocabulary (Gathercole, Frankish, Pickering & Peaker, 1999), and thus a cleaner way to measure the relation of WM to AR. We thus tested children's non-word memory span. To avoid floor effects, children who recalled no items were excluded.

3. Verbal inhibition: Another executive cognitive process that might cause perseverative AR errors is cognitive inhibition. The relation between inhibition and other executive functions on one hand, and AR and various other tests of response- or perspective-shifting on the other, is complex and controversial (e.g. Carlson & Moses, 2001; Hughes & Graham, 2002; Perner, 2000). We therefore tested a specific relation. If a word is activated strongly enough to be produced, it might interfere with production of other words. This interference is usually suppressed, but in very young children (i.e. 2-year-olds) verbal inhibition seems to work poorly, causing perseverative naming errors (Gershkoff-Stowe, 2002). AR errors might come from failures to inhibit the first AR response. If so, children with poor verbal inhibition ability should make many AR errors.

Despite the logical appeal of this hypothesis, it has no direct empirical support: Bialystok and Senman (2004)

found a mean partial r = .03 between AR scores and three tests of inhibition, with WM span controlled. Carlson and Moses (2001) found a mean partial r = .13 between AR scores and six tests of cognitive and response inhibition (controlling for age, gender and verbal ability). Carlson *et al.* (2002) found a mean partial correlation of r = .05 between AR and six inhibition tests, controlling for age and IQ. Thus, cognitive inhibition tests such as child-friendly Stroop tests do not predict children's AR responses. However, many of those tests do not focus on inhibition of lexical items (e.g. recently spoken words), which would be the kind of inhibition needed for correct AR performance. That is, the child must avoid saying the same word for both questions. Also, the inhibition process tapped by tests like the Stroop are confounded with working memory, making the results hard to interpret. A cleaner measure would be word intrusion errors such as those observed in verbal recall tests. Intrusion errors include repeating an item from a previous list, and producing high-frequency familiar words. The frequency of such errors might reflect children's difficulty with inhibiting lexical items. As such, intrusion errors might be closely related to perseverative AR errors. We therefore examined the relation of WM intrusion errors to AR errors.

4. Pragmatics: A very different explanation is that AR errors occur because children fail to understand adults' intended meaning – specifically, what the experimenter intends by asking the AR questions (Siegal, 1997). This idea is at the core of the *Pragmatics of Questions* account (Deák *et al.*, 2003), which stipulates that AR errors stem from a tendency to choose the same answer to several successive forced-choice questions. When preschoolers are asked several questions about an object, and must choose words to respond, they tend to choose the same word repeatedly. This peculiar error reveals a pragmatic failure: from question to question about a topic, listeners should focus on the new information in each question. Young children do not always favor new information when inferring message meanings (Deák, 2000; Luszcz & Bacharach, 1983). That is, whereas adults realize that different questions about a topic probably imply different aspects of the referent, preschool children do not. Focusing on the similarities between questions (e.g. identical response options), rather than differences, will cause perseveration. Thus, AR errors stem from a tendency to misinterpret forced-choice questions.

Only one study has tested the Pragmatics of Questions account. Deák *et al.* (2003) found that preschool children who perseverated in the AR test also perseverated in other control tests that posed successive forced-choice questions about the same item. The current study investigates the robustness of this tendency to perseverate across questions. A new *Forced Choice Question Pairs* (QP) task poses successive questions about two simple drawings on a single card (e.g. tree and flower). The questions each request a forced choice between words for the two pictures (e.g. 'Is the one that grows tall the tree or the flower?' and 'Is the one that smells good the tree or the flower?'). If the Pragmatics of Questions account is correct, children who perseverate in the AR task will also perseverate in the QP test. Because the QP test does not use deceptive objects or 'looks like'/'really' questions, accounts of AR errors based on specific conceptual limitations (Flavell *et al.*, 1983; Gopnik & Astington, 1988) or semantic/pragmatic ambiguities (Hansen & Markman, 2005) would not predict a specific relation between AR and QP test performance.

The Pragmatics of Questions account also suggests one reason *why* children perseverate in answering successive forced-choice questions: the emerging ability to detect whether a question is indeterminate (Deák *et al.*, 2003). Detecting indeterminacy is crucial for understanding many situations, for example, registering different possible locations for a hidden object (Fabricius, Sophian & Wellman, 1987), construing possible meanings of an ambiguous message (Revelle, Wellman & Karabenick, 1985), realizing that more than one agent might have caused an event (Sodian, Zaitchik & Carey, 1991) or seeing that an object might have come from any of several places (Fay & Klahr, 1996). These situations are indeterminate because more than one element could possibly fill the role of a variable in the problem (e.g. the agent of an event; Sodian *et al.*, 1991). Other cases, by contrast, allow only one specific value or entity to fill a variable. For example, if all items in a container are white, and an object is drawn from that container, then it must be white. This is a *determinate* problem. Distinguishing determinacy from indeterminacy is a complex cognitive skill. It has been described as distinguishing possibility and necessity (Piaget, 1987) or avoiding premature investment in one answer. The ability develops throughout childhood, so that even though children of 6 years or older still make errors (e.g. Speer, 1984), some preschool children show sensitivity to indeterminacy in simple situations (Fay & Klahr, 1996; Revelle *et al.*, 1985; Patterson, Cosgrove & O'Brien, 1980).

Variability in children's sensitivity to indeterminacy might explain perseverative AR errors. Adults expect different questions to focus on different aspects of a referent. Preschoolers, with sketchy pragmatic knowledge, do not have this expectation. If they do not notice the indeterminacy of each forced-choice question, but expect that *any* question about a topic has been answered (i.e. rendered determinate) once *one* question has been answered, they will tend to repeat the first

answer. Thus, faulty pragmatic expectations and a tendency not to notice the indeterminacy of specific questions about a topic might jointly cause children to repeat their first answer (i.e. perseverate).

Two results support this account. First, there is a correlation between children's failure to detect indeterminacy (e.g. claiming to know what color chip will be drawn next from a box of many-colored chips), and perseverating by inferring the same meaning for different words for an object (Deák & Narasimham, 2003). Second, children who perseverate in the AR and other successive forced-choice question tests are poor at detecting indeterminacy (Deák *et al.*, 2003). However, the latter finding was ambiguous because Deák *et al.*'s indeterminacy detection test used successive forced-choice questions (i.e. 'Do you know for sure . . . or do you have to guess?') about indeterminate and determinate scenarios, with similar wording and content. This test format might have been responsible for the correlation. In the current study we modified the Indeterminacy Detection (ID) test to eliminate this confound. There were no successive problems with similar questions or topics, and specific answer choices and referents changed in every trial. Thus, there is minimal basis for perseverating across questions. If ID and AR errors remain correlated, it will not be because procedural details elicited perseveration in the ID test.

Thus, the Pragmatics of Questions account makes two hypotheses. The first is descriptive: young children tend to perseverate in any task that poses successive, repetitive forced-choices questions about an object (i.e. topic). AR errors stem from a discourse context, not specific AR content (e.g. deceptive objects). The second is explanatory: children perseverate in part at least because they cannot tell indeterminate from determinate problems. After answering one question about a topic, they construe subsequent questions as already answered, or determinate. Thus, sensitivity to indeterminacy should predict AR performance. Note that this account does not rule out other causal contributors to perseveration (e.g. Hansen & Markman, 2005).

Summary

Several alternative accounts were tested by giving 3- and 4-year-old children an object-identity AR test and four others: *Flexible Naming* (FN), to assess children's capacity to flexibly activate and verbalize complex representations of objects; a *Working Memory* test of verbal span and of lexical inhibition (i.e. intrusion errors); *Question Pairs* (QP) to assess children's general tendency to perseverate across questions; and *Indeterminacy Detection* (ID) to test children's awareness of whether a situation has multiple possible answers or outcomes.

Method

Participants

Forty-two 3- and 4-year-olds (16 girls and 26 boys; mean age = 49 months, range 39–57) were tested. Two children who recalled no non-words in the WM test were replaced. The final sample included 17 3-year-olds (7 girls; mean = 44 months, range 39–47) and 25 4-year-olds (9 girls; mean = 53 months, range 49–57). Children were recruited from preschools in Southern and Northern California, and were mostly European-American and middle class.

Materials

Appearance–reality

Three deceptive objects were used: a banana-magnet, a lipstick-pen and a strawberry-eraser (see Deák *et al.*, 2003).

Flexible naming

Three representational objects were used (Deák & Maratsos, 1998): a dog puppet, a corncob-shaped pen and a car-shaped book. Most 3-year-olds produce function and appearance labels for each of these objects.

Forced-choice question pairs

Three 20×712 cm laminated cards each showed two clipart images: bird and dog, tree and flower, and boat and car.

Indeterminacy detection

The ID test problems used multicolored Lego blocks, white plastic rings, a box and a penny.

Procedure

Four tasks were given in one session in fixed order: AR, FN, WM and QP. ID scenarios were given between tasks and at the beginning and end of the session, in random order. Children were tested in a quiet room at their preschool. Sessions were videotaped.

Appearance–reality

As in Flavell *et al.* (1986), children saw deceptive objects one at a time (in random order). They identified (named) each object by its appearance, and its true function was then demonstrated and labeled. Thus, both words (e.g.

banana and *magnet*) were elicited before proceeding. Children then answered an appearance question and a reality question: 'When you look at this, does it look like a dog or look like a puppet?' and 'What is this really – is it really a dog or really a puppet?' Question and word order were counterbalanced.

Flexible naming

Children saw three representational objects, one at a time, in random order (Deák & Maratsos, 1998). Children were asked to name the object. The experimenter then elicited more labels by asking 'What else is it?', demonstrating the object's function, and asking about contrasting superordinate categories (e.g. 'What kind of thing is [a dog]? Is it a plant?'). Children then were asked to verify every possible pair of labels they had produced (e.g. 'Is it a dog and an animal at the same time?'), plus an equal number of foil pairs (e.g. 'Is it a cat and a puppet . . . ?') to check for inattention or a 'yes' response bias.

Follow-up questions: When a child produced at least one function and one appearance label for an object, they were asked, of each word, 'Is it called [word] because you use it like a ___, or because it looks like a ___?' Option and word order were randomized.

Verbal working memory

Children practiced a 'say after me' game with monosyllabic English words (two pairs, in separate trials), with feedback. They then heard three lists of four monosyllabic, pronounceable English non-words (some taken from Gathercole, Frankish *et al.*, 1999) in random order. The test non-words were *bon*, *chig*, *gub*, *kag*, *nish*, *nom*, *sab*, *shoon*, *tess*, *toop*, *parm* and *wark*. Children heard each list once and were asked to recall as many items as they could, without feedback. There was a 30-sec break between lists.

Forced-choice question pairs

Children were shown three cards, one at a time. They identified each picture on the card, then answered two questions in succession:

- '[Is the one that smells good . . .]/[Is the one that grows very tall . . .] a tree or a flower?'
- '[Is the one that goes on the water . . .]/[Is the one that has four wheels . . .] a boat or a car?'
- '[Is the one that is high in the sky . . .]/[Is the one that chews on bones . . .] a dog or a bird?'

Both questions about a given card ended with the same two word choices. Card, question and choice order all were randomized.

Indeterminacy detection test

Children did two training scenarios and six test scenarios. Each scenario implies either a single *determinate* answer or an *indeterminate* range of possible answers. Children judged whether the answer could be known with certainty, or not. Stimuli and wording, including the answer choices, differed for each scenario, to prevent perseveration due to repetitive questions. Scenario order was randomized, but both scenarios of one kind (e.g. color; location) were never given in sequence.

- *Training scenarios*: The determinate scenario was 'If I asked you what color my shirt is [pointing], would you have to guess the color, or could you tell for sure?' The indeterminate scenario was: 'If I asked you what I have in my pocket, would you know for sure, or would you have to guess?' Children received feedback and explanations. Scenario order was counterbalanced.
- *Color scenarios*: In the determinate scenario the experimenter visibly prepared to pull a ring from a box of identical white plastic rings, and asked 'Do you know what color [the ring I draw] will be, or will you have to guess what color it will be?' In the indeterminate scenario the box had different-colored Legos. Children were asked, '. . . if I get [one] out . . . do you know for sure what color it's going to be, or are you not sure what color it will be?'
- *Location scenarios*: In the determinate scenario the experimenter dramatically put a penny on his palm, closed his hand and asked, 'Can you tell me where [the penny] is, or do you have to look around to find it?' In the indeterminate scenario children were told, 'Yesterday I hid a marble somewhere in this big [building]. Could you get it for me right now, or would you have to look around for it?'
- *Meaning scenarios*: In the determinate scenario children were asked, 'Do you know what *blanket* is, or do you have to ask your teacher what a blanket is?' In the indeterminate scenario children were asked, 'What about the word *conifer*? Do you know what conifer is, or do you have to ask your mom what conifer is?'

Results

Coding

AR scores (range = 0–3) were one point per object for which a child chose the appearance word for the 'looks like' question *and* the function word for the 'really' question. FN scores (typically 1–3) were the mean number of appropriate words produced per object. FN follow-up

scores (range = 0–3) were the number of objects for which the child produced a function label and said it was for 'how it works', *and* produced an appearance label and said it was for 'what it looks like'. QP scores (range = 0–3) included one point for each card with both questions answered correctly. Perseverative errors also were coded in the AR, FN follow-up and QP tests. ID scores (range = 0–6) were one point per correct answer (i.e. determinate: 'know,' 'sure that' or 'can tell'; indeterminate: 'guess,' 'have to ask', 'don't know'). WM span (range = 0–12) was the total non-words recalled from all three lists (with 0.5 points for productions with only one incorrect phoneme). WM intrusion scores were the total number of incorrect English words or non-words from previous lists.

There were no gender differences in any measure except WM span: boys recalled more (mean = 3.2, SD = 1.2) than girls (2.4; 1.3), $t(40) = 2.2$, $p = .031$. Because this difference does not pertain to the current questions, we pooled boys' and girls' data in all further analyses.

Appearance–reality

Children chose correct words on a mean of 43% (SD = 40%) of trials. Twenty children (48%) perseverated on two or three trials. Most errors were repetitions of the function label (i.e. realism errors; see Figure 1), as is the norm (e.g. Flavell *et al.*, 1986). There were no significant item effects. AR scores increased with age, but not significantly, $r(40) = .24$, $p = .126$.

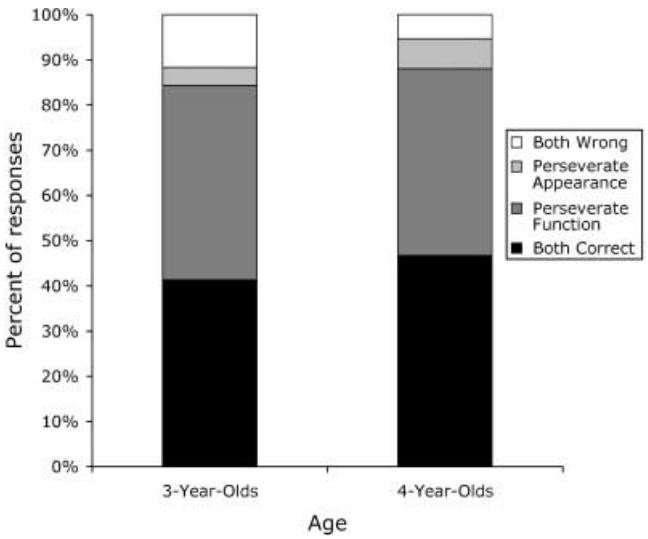

Figure 1 *Mean numbers of answer types in AR test: correct ('looks like' = appearance word; 'really is' = function word); perseveration on appearance word; perseveration on function word; both incorrect ('looks like' = function word; 'really is' = appearance word).*

Flexible naming

Children produced a mean of 2.5 words per object (SD = 0.6). Their age was not a significant predictor ($r = .10$). Children produced at least one appearance and one function word for 90% of objects. Most children (74%, or $n = 31$) did so for all objects. Children accepted a mean of 95% (SD = 11%) of correct label pairs (e.g. 'Is it a *dog* and a *puppet*?'), or 91% if we consider only those children who did not show a 'yes' response bias (i.e. by accepting many foil pairs). Thus, children uniformly accepted several labels for each object.

Children correctly answered 44% (SD = 33%) of follow-up questions about whether each word refers to the object's appearance or function. Three-year-olds perseverated on a mean of 47% of these pairs (e.g. said both words were for 'how you use it'); 4-year-olds perseverated on only 25%. This is a significant age difference, $t(40) = 2.3$, $p = .030$ (see Deák *et al.*, 2001).

Forced-choice question pairs

Children correctly answered a mean of 83% (SD = 30%) questions pairs. Twenty-nine children correctly answered all three pairs; 13 perseverated at least once. All errors were perseverative. Three-year-olds made marginally fewer correct answers (mean = 76%) than 4-year-olds (92%), $t(38) = 1.9$, $p = .062$.

Verbal working memory

Children recalled a mean of 3.4 (SD = 1.9) non-words, or 1.1 per list, similar to previous studies (Gathercole, Service *et al.*, 1999). Children made a mean of 2.4 intrusion errors (SD = 2.0). Most intrusions were similar-sounding familiar words; only four out of 99 were non-words from prior lists. Neither span nor number of intrusions differed between 3- and 4-year-olds, both ts$(40) < 1$.

Indeterminacy detection

Children averaged 4.0 correct (SD = 1.1) responses. The Lego-color problem was hardest ($n = 13$ correct); the Marble-location problem was easiest ($n = 40$ correct). ID scores were significantly correlated with age, $r(40) = .33$, $p = .033$. One question is whether the difficulty of a given scenario depended on the mental verb used. Location scenarios did not use typical mental or epistemic verbs but rather 'tell/look around' and 'get/look around', respectively, and received 28 and 40 correct responses. By contrast, scenarios with one epistemic verb, 'know', versus 'ask' or 'not sure' had 29, 33 and 13 correct

Table 1 *Top: Simple correlations among Appearance–reality (AR), Forced Choice Question Pair (QP), Indeterminacy Detection (ID), Flexible Naming (FN) words produced and follow-ups, and working memory (WM) span and intrusion errors. Bottom: Partial correlations with age, FN words produced and WM span removed*

Bivariate correlation	AR	QP	ID	FN words	FN follow-ups	WM span	WM intrusions
Age	.24	−.10	.33*	.10	.14	.06	−.04
AR		.36*	.44*	.24	.36*	.08	.12
QP			.43*	.12	.15	.23	.06
ID				.26	.37*	.17	.06
FN words					.11	.27	.09
FN follow-ups						.24	−.19
WM span							−.20

Partial correlation	AR	QP	ID	FN follow-ups	WM intrusions
AR		.39*	.36*	.33*	.13
QP			.50*	.11	.05
ID				.35*	.15
FN follow-ups					.03

Note: * $p < .05$; ** $p < .005$.

responses. Finally, the determinate color scenario had two epistemic verbs ('know/guess') and yielded 26 correct responses. Thus, three items that respectively had 0, 1 and 2 epistemic verbs received 28, 29 and 26 correct responses. This is not a compelling relation. However, 27 children got both location scenarios correct, whereas only seven got both color scenarios correct, suggesting a content effect.

Perseverative errors

Summed across the AR, FN follow-up and QP tests, children made a mean total of 2.9 perseverative errors (SD = 1.7) out of nine possible errors. The difference between 3-year-olds (3.0) and 4-year-olds (2.8) was not significant, $t(40) < 1$. As a stricter measure of perseveration, we also considered only those repetitions in which the first answer was correct. Children averaged 1.6 (SD = 1.2) of these 'first-correct' errors.

Relations between tasks

Simple correlations among age, AR, FN, FN follow-up, QP, ID, WM span and ID scores are shown in Table 1. AR, FN follow-up, QP and ID scores all were significantly correlated (except FN follow-up and QP scores), as predicted by the Pragmatics of Questions account. FN production was not correlated with AR, contrary to the representational inflexibility account. Neither WM span nor intrusions was correlated with AR scores, contrary to the memory span and inhibition accounts, respectively.

The bottom panel of Table 1 shows partial correlations between test. Age, WM span and FN total production were partialled out (the last is correlated with vocabulary; Deák & Maratsos, 1998). AR, FN follow-up, QP and ID scores all remained significantly correlated, except FN follow-up and QP scores. Thus, children who perseverated in the AR test also perseverated in other tests with successive forced-choice questions, and had trouble distinguishing determinate from indeterminate situations. However, virtually every preschooler, including those who perseverated in the AR test, could flexibly label different aspects of an object, including its appearance and function.

To better understand these relations, a linear stepwise regression was done on AR-combined scores; that is, number of AR items correct plus number of correct FN follow-up items (i.e. the alternate AR measure). Age, FN words produced, QP scores, ID scores, WM span and WM intrusions were entered as predictors. Only Indeterminacy Detection predicted AR performance: $\beta_{standardized} = .50$, $R^2_{adjusted} = .23$, $F(1, 40) = 13.2$, $p < .001$. No other factor explained additional variance. The results were the same using simple AR test scores as the dependent variable. Ability to detect indeterminacy predicted about one-quarter of variance in AR scores: far more, for example, than tests of false belief or working memory (Carlson & Moses, 2001).

The Pragmatics of Questions account also predicts that indeterminacy detection will determine QP scores. Another stepwise regression was done on QP scores with all other variables entered as factors. Only ID was a

significant predictor: $\beta_{standardized} = .42$, $R^2 = .18$, $F(1, 40) = 8.8$, $p = .005$. Age did not predict significant additional variance, nor did WM span, intrusions or FN production.

What predicted Indeterminacy Detection? When all factors except AR scores were entered in a stepwise regression, only age predicted ID scores ($\beta_{standardized} = .33$; $R^2_{adjusted} = .09$, $F(1, 40) = 4.9$, $p < .033$). Thus, awareness of indeterminacy improves with age. Presumably older children are more likely to recognize similar but distinct questions as possibly having different answers.

Importantly, all regression results are similar if, rather than using AR and QP scores as dependent measures, we use perseverative errors. This is not surprising, as most errors were perseverative. Further, we examined the sum of first-correct perseverative errors (from AR, QP and FN follow-up tests) in a stepwise regression, with age, ID scores, FN productivity, WM recall and intrusions entered. Only ID predicted first-correct perseveration: $\beta_{standardized} = -.53$, $R^2_{adjusted} = .26$, $F(1, 40) = 15.5$, $p < .001$. No other factor explained significant additional variance. This confirms that awareness of indeterminacy was the only factor to predict perseveration.

Because QP scores were skewed, we confirmed their relation to AR and ID scores by classifying each child as a QP perseverater (at least one error, $n = 13$) or non-perseverater (no errors; $n = 29$). Non-perseveraters had significantly higher AR-combined scores (difference = 1.1, $t(34) = 2.1$, $p = .046$) and higher ID scores (difference = 1.0, $t(26) = 3.0$, $p = .005$).

General discussion

Appearance–reality errors are akin to perseverative errors made in other tests that have a similar questioning format. According to the Pragmatics of Questions account, preschoolers might not realize that when asked several questions about a topic, each question probably implies a different answer. Children who know this should not repeat an answer for several different questions, because they register each question as a new indeterminacy that requires processing of the content and the response options. Although the current study does not prove this hypothesis, it confirms findings (Deák *et al.*, 2003) that children's sensitivity to indeterminate situations is negatively related to perseverative errors. The association is robust, even when age, memory span and tendency to make verbal intrusions are controlled.

A critical finding not predicted by other accounts is that perseveration was correlated across tests with successive forced-choice questions. This indicates that questioning format, rather than specific AR content (e.g. deceptive objects), causes children's perseverative question-answering errors, including most AR-test errors. An exception is that FN follow-up and QP scores were not reliably related, but this might be because FN follow-up scores depended partly on FN productivity, or because follow-up questions involve choices between complex phrases ('what it looks like or what you do with it') rather than simple words as in the QP test. Regardless, perseveration in all of these tests was predicted by indeterminacy detection.

The results also confirm previous findings that AR errors are not due to representational inflexibility (Deák & Maratsos, 1998; Deák *et al.*, 2001, 2003; Hansen & Markman, 2005; Sapp *et al.*, 2000). For example, all children produced appearance and function words for objects in the FN test, but many of the same children made AR errors. Perhaps, however, a *specific* kind of representational inflexibility causes AR errors. Some theorists have claimed that the capacity to represent false or changed beliefs (e.g. about an object's identity) determines AR ability. This was implied by Gopnik and Astington's (1988) report that scores on AR and False Belief (FB) tests share about 20% of variance. However, that study did not control for age, memory or verbal skills, and used identical stimuli and scenarios in all tasks. Thus, the finding is not interpretable. More recent studies have found small or null correlation between AR and FB tests (e.g. Carlson *et al.*, 2002; Miller, Holmes, Gitten & Danbury, 1997), suggesting the first report was spurious. Even if AR and FB scores are correlated, the Pragmatics of Questions account would predict a correlation anytime both tests use a similar question format (e.g. successive questions with the same forced choices, as in Gopnik & Astington, 1988). Differences in AR–FB correlations across experiments might be due to procedural differences, for example, using successive forced-choice questions in one test but not the other. For example, Carlson *et al.* (2002) used successive forced-choice questions in the AR test and in one FB trial, but not the other, which might explain their low AR–FB correlation. The possibility that discourse factors determine whether AR and FB tests will correlate suggests that the former should not be used as a measure of theory-of-mind development.

The Pragmatics of Questions account also can explain findings that AR accuracy improves when the discourse structure of the task is altered. For example, asking only one forced-choice question about an object helps (Hansen & Markman, 2005). Also, children perseverate less when choosing between objects instead of words (Rice *et al.*, 1997; Sapp *et al.*, 2000), possibly because choosing objects highlights the task's indeterminacy by making alternative choices concrete.

Hansen and Markman (2005) suggest a different pragmatics-based account of AR errors. They note that 'looks like . . .' can mean either 'is' or 'resembles',

and without additional pragmatic cues children tend to choose the former, reality-based interpretation. This leads them to choose the function word for both AR questions. This account makes important points: preschool children *do* find 'looks like' ambiguous (Deák, 2000), and their dominant error is to choose the function label twice (i.e. realism errors). However, it cannot explain other findings. First, children sometimes perseverate on the appearance word rather than the function word (see Figure 1). Second, Deák *et al.* (2003) compared the standard ambiguous AR questions to less ambiguous questions, and found no difference in errors (see also Flavell, Green, Wahl & Flavell, 1987). Third, AR and QP errors are correlated, and ID scores predicted both of these. Neither the QP nor the ID test uses 'looks like' questions, so Hansen and Markman's account would not predict these correlations. Deák *et al.* (2003) found that AR errors correlate with perseveration in another control task that did not use 'looks like' questions. Thus, Hansen and Markman's account is at best incomplete, though it is correct insofar as young children *can* assign items to multiple categories and label them as such, and 'looks like' is ambiguous to children. The latter might explain why we found a higher error rate in the AR than in the QP test. That is, beyond any general tendency to perseverate, AR questions are relatively hard, and this might be due to the ambiguity of 'looks like'. In sum, Hansen and Markman's account does not explain all the same results as the Pragmatics of Questions account, but it might explain why the AR test is fairly hard. Note that the accounts are neither contradictory nor mutually exclusive.

The results are not readily explained in terms of development of two executive functions, working memory and verbal inhibition. Accounts focusing on these factors (Bialystok & Senman, 2004; Hughes & Graham, 2002; Perner, 2000) are plausible because both improve in early childhood (e.g. Gathercole & Adams, 1993; Luciana & Nelson, 1998), during the period when AR performance improves. Yet there is no support for a close relation between AR errors and either WM or inhibitory capacity. Several findings show a weak relation between memory span and AR errors (Carlson *et al.*, 2002; Deák *et al.*, 2003). The current results corroborate the non-significant relation using a non-word span test that is less reliant on lexical knowledge than word or digit span tests. Also, the current results avoid a floor effect that might have contributed to Deák *et al.*'s (2003) negative result. Thus, although the AR test seems to have working memory demands, there is no evidence that this contributes to preschoolers' AR errors. Also, there seem to be weak or null relations between response inhibition and AR errors (Carlson & Moses, 2001; Carlson *et al.*, 2002). Lexical intrusions (which suggest poor verbal inhibition) are not correlated with perseverative AR errors. Intrusions are a good measure because, unlike other putative inhibition tests (e.g. Stroop-like tests; Carlson *et al.*, 2002), they do not conflate inhibition with memory for task rules. Also, because AR errors can be viewed as verbal intrusions, a test of other verbal intrusions would seem to address a fairly specific inhibition-based account. However, the results support the general claim that perseverative errors are not always due to inhibitory failure (Deák & Narasimham, 2003).

In sum, the Pragmatics of Questions account explains a wide range of AR results, whereas accounts based on representational flexibility, working memory or lexical inhibition do not. However, two other accounts should be considered. One would explain AR errors in terms of developing *epistemic knowledge*: that is, conceptual knowledge about knowledge states (e.g. guessing vs. knowing; O'Neill & Gopnik, 1991). Children who can explicitly reflect on their own epistemic states should be able to represent the change in their belief about an object's identity, between first seeing it and then seeing its function demonstrated. They would then use the memory of their own belief change, based on different kinds of information, to choose the answers to each AR question. This could explain the current results, including the correlation with ID scores, which might also depend on the ability to reflect on one's own knowledge states. This account is, however, less parsimonious than the Pragmatics of Questions account, because it adds the process of *reflecting* on and/or labeling one's epistemic states. Pragmatics of Questions does not require an explicit reflection process. It merely suggests that preschoolers vary in how well they generate an uncertainty 'signal' in indeterminate situations, and use it to choose overt responses (e.g. between the options 'sure' and 'not sure'). Also, an epistemic knowledge account cannot explain the association of AR and ID errors with QP errors. Finally, the number of epistemic verbs in each scenario did not strongly predict item effects. Instead, location scenarios were easier than color scenarios. This could be due to experience: all children have searched for misplaced objects and wrongly guessed about their locations, whereas guessing a hypothetical object's color might be an unfamiliar experience. This speculation highlights the need for future research on factors that contribute to children's inferences about indeterminacy. In sum, epistemic knowledge might contribute to ID and AR performance, but it does not explain all results, and it is not as parsimonious as the Pragmatics of Questions account.

Second, the Alternative Perspectives account (Doherty, 2000; Perner, Strummer, Sprung & Doherty, 2002) suggests

that children have trouble coordinating two veridical perspectives on an object or event. Thus, children who fail a false belief task also fail to produce a synonym or superordinate category label for a just-named item. There is some support for the Alternative Perspectives account (Perner *et al.*, 2002). In addition, it would predict some of the present results, like the AR–ID correlation, and fewer QP errors. However, it is not consistent with evidence that 3-year-olds readily produce several labels in the Flexible Naming test and confirm label pairs (e.g. '. . . a dog *and* animal'). Also, it does not explain the AR and QP correlation, because the latter does not require coordinating two perspectives. In other regards, the Alternative Perspectives and Pragmatics of Questions accounts are compatible. If we postulate that detecting indeterminacy helps children confront different perspectives on a complex topic, the two accounts can be synthesized, and thereby explain most findings on the object-identity AR test and other tests as well (Perner *et al.*, 2002). Future work should explore whether detecting indeterminacy helps children represent multiple perspectives, and thus correctly answer several questions about a topic.

In sum, the current findings help narrow down possible causes of appearance–reality errors. Other questions remain, however. For example, do the data generalize from object-identity AR tests to property AR tests (e.g. apparent vs. real color)? Why are the AR and FN follow-up tests harder than the QP test (e.g. nature of the forced choices)? How does indeterminacy detection develop? Finally, what factors account for the unexplained variance in the AR test? Future studies should include comprehensive language assessment tests, as well as tests of epistemic knowledge and executive functions. The discourse format of all tests must be carefully controlled. Finally, the ID test should be refined to include nonverbal tests of sensitivity to indeterminacy (Revelle *et al.*, 1985). This would bring us closer to fully explaining children's appearance–reality errors.

Acknowledgements

Brian Enright submitted a version of this study as an honors thesis to the Cognitive Science Department at UCSD. We thank the children who participated, and Jeff Elman, Cristine Legare and Leah Welch for helpful suggestions.

References

Andrews, G., Halford, G.S., Bunch, K., Bowden, D., & Jones, T. (2003). Theory of mind and relational complexity. *Child Development*, **74**, 1476–1499.

Bailystok, E., & Senman, L. (2004). Executive processes in appearance–reality tasks: the role of inhibition of attention and symbolic representation. *Child Development*, **75**, 562–579.

Carlson, S.M., & Moses, L.J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, **72**, 1032–1053.

Carlson, S.M., Moses, L.J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant and Child Development*, **11**, 73–92.

Clark, E.V., & Svaib, T.A. (1997). Speaker perspective and reference in young children. *First Language*, **17**, 57–74.

Deák, G.O. (2000). The growth of flexible problem solving: preschool children use changing verbal cues to infer multiple word meanings. *Journal of Cognition and Development*, **1**, 157–192.

Deák, G., & Maratsos, M. (1998). On having complex representations of things: preschoolers use multiple words for objects and people. *Developmental Psychology*, **34**, 224–240.

Deák, G.O., & Narasimham, G. (2003). Is perseveration caused by inhibition failure? Evidence from preschool children's inferences about word meanings. *Journal of Experimental Child Psychology*, **86**, 194–222.

Deák, G.O., Ray, S.D., & Brenneman, K. (2003). Children's perseverative appearance–reality errors are related to emerging language skills. *Child Development*, **74**, 944–964.

Deák, G.O., Yen, L., & Pettit, J. (2001). By any other name: when will preschoolers produce multiple labels for a referent? *Journal of Child Language*, **28**, 787–804.

Doherty, M.J. (2000). Children's understanding of homonymy: metalinguistic awareness and false belief. *Journal of Child Language*, **27**, 367–392.

Fabricius, W.V., Sophian, C., & Wellman, H.M. (1987). Young children's sensitivity to logical necessity in their inferential search behavior. *Child Development*, **58**, 409–423.

Fay, A.L., & Klahr, D. (1996). Knowing about guessing and guessing about knowing: preschoolers' understanding of indeterminacy. *Child Development*, **67**, 689–716.

Flavell, J.H., Flavell, E.R., & Green, F.L. (1983). Development of the appearance–reality distinction. *Cognitive Psychology*, **15**, 95–120.

Flavell, J.H., Green, F.L., & Flavell, E.R. (1986). Development of knowledge about the appearance–reality distinction. *Monographs of the Society for Research on Child Development*, **51** (1, serial no. 212).

Flavell, J.H., Green, F.L., Wahl, K.E., & Flavell, E.R. (1987). The effects of question clarification and memory aids on young children's performance on appearance–reality tasks. *Cognitive Development*, **2**, 127–144.

Frye, D., Zelazo, P.D., & Palfai, T. (1996). Theory of mind and rule-based reasoning. *Cognitive Development*, **10**, 483–527.

Gathercole, S.E., & Adams, A. (1993). Phonological working memory in very young children. *Developmental Psychology*, **29**, 770–778.

Gathercole, S.E., Frankish, C.R., Pickering, S.J., & Peaker, S.H. (1999). Phonotactic influences on serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **25**, 84–95.

Gathercole, S.E., Service, E., Hitch, G.J., Adams, A., & Martin, A.J. (1999). Phonological short-term memory and vocabulary development: further evidence on the nature of the relationship. *Applied Cognitive Psychology*, **13**, 65–77.

Gershkoff-Stowe, L. (2002). Object naming, vocabulary growth, and the development of word retrieval abilities. *Journal of Memory and Language*, **46**, 665–687.

Gopnik, A., & Astington, J.W. (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance–reality distinction. *Child Development*, **59**, 26–37.

Hansen, M.B., & Markman, E.M. (2005). Appearance questions can be misleading: a discourse-based account of the appearance–reality problem. *Cognitive Psychology*, **50**, 233–263.

Hughes, C., & Graham, A. (2002). Measuring executive functions in childhood: problems and solutions? *Child and Adolescent Mental Health*, **7**, 131–142.

Luciana, M., & Nelson, C.A. (1998). The functional emergence of prefrontally-guided working memory systems in four- to eight-year-old children. *Neuropsychologia*, **36**, 273–293.

Luszcz, M.A., & Bacharach, V.R. (1983). The emergence of communicative competence: detection of conversational topics. *Journal of Child Language*, **10**, 623–637.

Melot, A.-M., & Houdé, O. (1998). Categorization and theories of mind: the case of the appearance/reality distinction. *Cahiers de Psychologie Cognitive/Current Psychology of Cognition*, **17**, 71–93.

Miller, S.A., Holmes, H.A., Gitten, J., & Danbury, J. (1997). Children's understanding of false beliefs that result from developmental misconceptions. *Cognitive Development*, **12**, 21–51.

O'Neill, D.K., & Gopnik, A. (1991). Children's ability to identify the sources of their beliefs. *Developmental Psychology*, **27**, 390–397.

Patterson, C.J., Cosgrove, J.M., & O'Brien, R.G. (1980). Nonverbal indicants of comprehension and noncomprehension in children. *Developmental Psychology*, **16**, 38–48.

Perner, J. (2000). About + belief + counterfactual. In P. Mitchell & K. Riggs (Eds.), *Children's reasoning about the mind* (pp. 367–401). Hove, UK: Psychology Press.

Perner, J., Strummer, S., Sprung, M., & Doherty, M. (2002). Theory of mind finds its Piagetian perspective: why alternative naming comes with understanding belief. *Cognitive Development*, **17**, 1451–1472.

Piaget, J. (1987) *Possibility and necessity* (trans. H. Feider). Minneapolis, MN: University of Minnesota Press.

Revelle, G.L., Wellman, H.M., & Karabenick, J.D. (1985). Comprehension monitoring in preschool children. *Child Development*, **56**, 654–663.

Rice, C., Koinis, D., Sullivan, K., Tager-Flusberg, H., & Winner, E. (1997). When 3-year-olds pass the appearance–reality test. *Developmental Psychology*, **33**, 54–61.

Sapp, F., Lee, K., & Muir, D. (2000). Three-year-olds' difficulty with the appearance–reality distinction: is it real or is it apparent? *Developmental Psychology*, **36**, 547–560.

Siegal, M. (1997). *Knowing children: Experiments in conversation and cognition* (2nd edn.). Hove, UK: Psychology Press.

Sigel, I.E., Saltz, E., & Roskind, W. (1967). Variables determining concept conservation in children. *Journal of Experimental Psychology*, **74**, 471–475.

Sodian, B., Zaitchik, D., & Carey, S. (1991). Young children's differentiation of hypothetical beliefs from evidence. *Child Development*, **62**, 753–766.

Speer, J.R. (1984). Two practical strategies young children use to interpret vague instructions. *Child Development*, **55**, 1811–1819.

Received: 5 July 2004
Accepted: 13 September 2005