Child Development, May/June 2003, Volume 74, Number 3, Pages 944–964

Children's Perseverative Appearance–Reality Errors Are Related to Emerging Language Skills

Gedeon O. Deák, Shanna D. Ray, and Kimberly Brenneman

Two experiments explored the communicative bases of preschoolers' object appearance–reality (AR) errors. In Experiment 1, 3-, 4-, and 5-year-olds ($N = 36$) completed the AR test (with high- and low-deceptive objects), a control test with the same discourse structure but nondeceptive stimuli, and stimulus naming and memory tests. AR performance correlated positively with control (discourse) and naming test performance. Object deceptiveness had little effect. In Experiment 2, 3- and 4-year-olds ($N = 64$) completed AR tests that experimentally varied question phrasing and use of exemplar objects. Children also completed memory, vocabulary, and control tests (of verbal perseveration). AR performance variance was predicted by a composite perseveration score from three non-AR tasks, vocabulary, and exemplars. The results indicate that the discourse structure of the AR test elicits a perseverative tendency that is mediated by children's verbal knowledge.

Adults in modern societies are accustomed to illusion. Our surroundings feature disembodied voices floating from stereo speakers, colored light on TV screens showing fantastic creatures and events, refrigerator magnets resembling juicy victuals or cute animals, and magic tricks that transform ordinary objects. Adults are entertained but not fooled by such phenomena; it is less clear how young children understand them. What are the sources of children's erroneous answers to questions about apparent and real aspects of deceptive objects—items that look like one thing but function like another? Such errors often are assumed to stem from an inability to represent two concepts at once, or an inability to represent one's own changing beliefs about an object's identity. We explore an alternative account: Children's errors might be due to specific communicative and linguistic processes. By this account, mature responses to questions about appearance and reality require the respondent to map each question (or its semantic content) to alternate descriptions of the deceptive stimulus, and to understand the discourse structure of the task. Children who lack these capacities will tend to err, usually by giving the same answer to successive, different questions. The studies reported here are intended to assess this account of appearance–reality (AR) errors.

Naming Deceptive Objects: Questions and Answers

Children's ability to distinguish between appearance and reality was first systematically studied by Flavell and colleagues (e.g., Flavell, Flavell, & Green, 1983, 1987; Flavell, Green, & Flavell, 1986). Though other researchers have devised complementary paradigms (DeVries, 1969; Harris, Donnelly, Guz, & Pitt-Watson, 1986; Wellman & Estes, 1986), Flavell's tasks have become standards for testing children's grasp of the AR distinction. One, the object identity task, focuses on objects that look like one thing but are really another, such as a magnet that looks like a tomato or a candle that resembles a rock. Such deceptive objects are a type of representational object (Deák & Maratsos, 1998), that is, artifacts made to resemble a specific kind or individual. Deceptive objects require focused or extraordinary examination to discover the discrepancy between what they are meant to resemble and how they are meant to function.

In object identity tasks, children name a deceptive object before its true function is revealed. Children

Gedeon O. Deák, Department of Cognitive Science, University of California–San Diego; Shanna Ray, Department of Psychology, David Lipscomb University; Kimberly Brenneman, Department of Psychology, the State University of New Jersey, Rutgers.

Thanks to the children, parents, and teachers at Eakin Elementary School extended care program, and at the following preschools (Nashville, Tennessee): Bellevue Presbyterian, Belmont United Methodist, Brentwood Methodist, Cooperative Child Care, Hillwood Presbyterian, Harpeth Heights Baptist, Vanderbilt, Woodmont Christian, and Woodmont Hills Church of Christ. Thanks to Kate Adams and Anne Wolff for assistance in collecting and coding data. Thanks also to three anonymous reviewers for helpful suggestions. The research was supported by the Vanderbilt University Research Council. Some results were presented at the 28th Annual Jean Piaget Society Symposium, Chicago, 1998.

Correspondence concerning this article should be addressed to Gedeon Deák, Department of Cognitive Science, 9500 Gilman Drive, University of California–San Diego, La Jolla, CA 92093-0515. Electronic mail may be sent to deak@cogsci.ucsd.edu.

then are asked two successive questions, one focusing on appearance and the other on reality (defined by function). Each question poses a choice between two labels, one denoting the object's apparent kind, the other denoting its function. For instance, after learning that a rock-looking object is really a sponge, a child is asked ''What does this look like, a rock or a sponge?'' and ''What is this really and truly, a rock or a sponge?'' (Flavell et al., 1986). It is deemed correct (i.e., adultlike) to answer the ''looks like'' question with the appearance-kind label (i.e., rock), and to answer the ''really and truly'' question with the function-kind label (i.e., sponge). Other answers are deemed incorrect: Saying the appearance label twice is deemed a phenomenism error because it seems to focus on the perceptually salient kind; saying the function label twice is a realism error because it seems to focus on the ''true'' kind. Children seldom switch the two labels.

Children's responses show a predictable developmental function. Five-year-olds usually answer standard object AR questions correctly. Three-year-olds produce many realism errors, fewer phenomenism errors or correct responses, and very few switches. Four-year-olds typically produce 40% to 60% errors, mostly of the realist type (Brenneman & Gelman, 1993; Flavell et al., 1983; Gauvain & Greene, 1994; Gopnik & Astington, 1988; Ray, 1996; Taylor & Hort, 1990). Random responding would yield 75% errors. Thus, few 3- to 4-year-olds consistently and appropriately shift between valid labels for deceptive objects.

Explaining Children's AR Errors

It is often assumed that AR errors stem from difficulty representing mental states, particularly false or changed beliefs. This assumption rests on evidence that AR, false belief, and perspective taking test performance all improve roughly around the same age (Flavell, Green, & Flavell, 1989; Frye, 2000; Gopnik & Astington, 1988; Harris & Leevers, 2000). The dearth of additional evidence for this assumption, however, suggests the need to seek alternative accounts. The most prominent alternative is that AR errors stem from representational inertia: some difficulty maintaining, coordinating, or switching between multiple representations of a situation or stimulus (e.g., Flavell et al., 1986). The representational inertia account, however, seems incompatible with evidence that 3-year-olds can label multiple aspects of representational objects, from experimental tasks similar to the AR paradigm (Clark & Svaib, 1997; Deák & Maratsos, 1998; Deák, Yen, & Pettit, 2001). For example, Deák and Maratsos (1998) asked preschool children several questions about representational objects such as a puppet dog. The questions were designed to elicit labels for both the appearance and the function of each object, and even 3-year-olds frequently responded by producing two or more words for an object. If representational inertia causes AR errors, it is not clear how 3-year-olds can shift labels so readily in Deák and Maratsos's task.

A limited, specific version of the representational inertia account remains plausible in the face of such evidence. Specifically, inertia might be due to limited working memory capacity, so that even if young children can conceptually grasp two categorical representations of an entity, they cannot produce both labels under challenging conditions such as the absence of perceptual evidence for both categories. This condition is manifest in deceptive objects. If this account is correct, children should be aided by perceptual support that reduces memory demands. Consistent with this prediction, Rice, Koinis, Sullivan, Tager-Flusberg, and Winner (1997) and Brenneman and Gelman (1993) found that presenting exemplars of both categories of an AR object (e.g., a real apple and a real candle, for a deceptive apple-candle) improved children's AR performance. One explanation is that the exemplars reduced working memory demands of the AR task (Rice et al., 1997). For example, Deák and Maratsos (1998) found that 3- and 4-year-olds named both appearance and function of representational but nondeceptive objects, in which both appearance and function were perceptually apparent (e.g., the puppet dog clearly was meant to represent a dog but was obviously a puppet). Perhaps the unambiguous perceptual cues helped children maintain both categories and labels in working memory. If perceptual evidence mediates young children's AR performance by affecting memory demands of the task, their performance should be affected by object deceptiveness. That is, truly deceptive objects have abundant perceptual cues to the appearance category but few readily available cues to the function category. This asymmetry in perceptual cues to the two categories, and category labels, might dispose children with limited verbal working memory to perseverate when labeling deceptive objects but not when labeling nondeceptive objects such as Deák and Maratsos's stimuli. To test this hypothesis, Experiment 1 tested children's response to high- and low-deceptive AR objects.

The working-memory-based representational inertia account also makes predictions about the

source of individual differences in AR performance. The AR task imposes a verbal working memory load that is nontrivial for young children. When each question is posed the child must represent the predicate (e.g., "looks like") and two response options—essentially a two-word list—to evaluate each word's appropriateness with respect to the predicate. Preschool children's immediate memory span is two to four words (Gathercole & Adams, 1993); therefore, the AR task might approach the limits of children's verbal working memory. Because both AR performance and verbal working memory span vary across 3- to 5-year-olds, the two measures might be related. To test the relation between AR performance and working memory capacity, children in Experiments 1 and 2 completed an immediate word recall test.

Another explanation of the development of AR test performance involves changes in language skills. These skills are critical because the AR test requires choosing lexical responses to semantically distinct questions. Few studies, however, have focused on the linguistic demands of the AR test. Perhaps this dearth of evidence is due to early reports that no improvement was found in AR performance when either the standard questions were modified (Flavell et al., 1986; Flavell, Green, Wahl, & Flavell, 1987) or children received brief training on the test questions (Taylor & Hort, 1990). However, these studies had low statistical power and provided minimal training on the meanings of the questions or the pragmatic implications of the test questions. Also, these studies did not show that the modified questions were any clearer than the standard questions.

It is therefore significant that a recent investigation showed that verbal demands of the AR test contribute substantially to children's errors (Sapp, Lee, & Muir, 2000). Three-year-olds correctly solved nonverbal object AR problems (i.e., choosing between objects) but made many errors on verbal problems (i.e., choosing between words). We still do not know, however, which verbal factors affect preschoolers' performance. One clue is that AR errors are fundamentally perseverative; that is, they are inappropriate repetitions of prior naming responses. The AR test poses the same response options (i.e., labels) for two successive questions. To answer correctly, children must grasp the specific implications of each question with respect to the object labels. Such tasks seem to elicit perseveration across a fairly wide range of stimuli or questions. For example, Deák (2000) asked 3- to 6-year-old children several questions about each of several novel objects. Each question about a given object featured a different novel word, following a unique predicate (e.g., "is made of …," "has a …") which implied some property (material, part, or shape). Children had to infer the meaning of each novel word and to generalize it to another object that shared a critical property, based on the predicate. Inappropriate responses (i.e., generalizations that were unrelated to the predicate) were mostly perseverative: Three- and 4-year-olds tended to assign the same meaning to several words, though the question predicate had changed. By analogy, AR predicates ("is really and truly," "looks like") should imply distinct meanings, but 3- and 4-year-olds respond as if they do not. In both of these tests, then, 3- and 4-year-old children tend to perseverate across distinct questions about complex stimuli.

These findings suggest that AR errors are just one example of perseveration in label selection or production. But why do preschoolers perseverate in naming objects? The typical explanation for perseveration is that the subject could not inhibit a dominant or highly activated response (Dempster, 1992; Houdé, 2000). Another explanation is that younger children are unable to control their selection of responses when the task contingencies that dictate these choices become too complex (Frye, Zelazo, & Palfai, 1995), and in these cases children maintain previously selected responses even when the contingencies shift. A third explanation is that children stick to a highly confident first response if they do not recognize that later problems are substantially different from previous problems (Deák, 2000). None of these accounts specifically focuses on the role of language abilities in perseveration, but Sapp et al.'s (2000) data suggest that linguistic demands are an important element of AR perseveration. In contrast, a fourth explanation focuses on pragmatic aspects of perseverative errors, with regard to AR and false belief tasks. Siegal (1991) attributed errors in these tasks to a mismatch between younger children's pragmatic knowledge and the tasks' violation of conversational patterns familiar to children. This is an important consideration, but it does not explain why children's errors are perseverative. If, that is, children perceive the adult's questions as silly, why do they perseverate instead of, for instance, making playful or idiosyncratic responses (i.e., playing with the adult) or requesting clarification?

We consider a slightly different account of AR errors based on children's discourse and lexical knowledge. The AR task demands analysis of every question's specific, distinct meaning. Adults recognize this demand when answering questions, but

awareness of the demand might follow a long process of learning the conventions of questioning in formal educational settings (i.e., schools). Informal observation suggests that 2- and 3-year-olds are motivated not to understand questions and give correct answers, but rather to interact with adults without eliciting negative affect or corrective feedback (see Donaldson, 1978; Siegal, 1991). When adults seem happy, the child perceives the situation as all okay and has no motive to change answers across questions.

Furthermore, we assert that preschool children grasp the conceptual dissociation in deceptive objects and can produce multiple labels (Deák & Maratsos, 1998). However, 3- and 4-year-old children are insensitive to the meanings of questions that imply specific labels. Sensitivity to distinct meaning is crucial in the AR test, where each question's intended meaning is conveyed principally by predicates (e.g., ''is really and truly a …''). Perhaps 3-year-olds, for some reason, tend not to notice changes in meaning across questions. In keeping with this idea, Olson (1977) argued that young children do not decontextualize messages to interpret them (i.e., do not focus strictly on the literal semantic content), a practice tightly tied to literacy and formal education. Thus, when choosing verbal responses (e.g., labels), rather than focusing on the semantics of the current question, preschoolers might homogenize their interpretation of several recent questions. Homogenization of meanings could cause perseveration because successive questions are not interpreted as distinct.

The idea that perseveration in the AR task stems from blending the meanings of different questions goes beyond Siegal's (1991) claim that the AR test violates pragmatic conventions. That is, preschool children not only perceive the experimenter's questions as silly or puzzling but as indistinct or undifferentiated. A unique prediction of this account is that deceptiveness of AR objects and specific phrasing of standard AR questions are not central to children's errors. Errors might reveal a general tendency to perseverate across questions, at least in some discourse conditions. To determine whether discourse conditions, rather than deceptive objects or specific questions, elicit AR errors, in Experiments 1 and 2 we administered control tests with the same discourse structure as the AR task, but with nondeceptive objects and different questions.

A second prediction is that some verbal knowledge, for example, depth and breadth of word meaning knowledge, mediate children's errors in the AR test. In Deák and Maratsos (1998) and Deák et al. (2001), receptive vocabulary—a rough index of lexical knowledge—predicted the number of words children produced for representational objects. In the AR test, children's ability to produce two labels for an object (i.e., correct responses) might depend on knowledge of word meanings. For example, if children do not fully comprehend both labels for a deceptive object (e.g., apple and candle), they might perseverate on one or the other. A similar account of children's errors was suggested by Merriman, Jarvis, and Marazita (1995). Scrutiny of AR stimulus object labels in several studies lends credence to the idea that understanding both labels influences AR performance. Although some labels (e.g., crayon and apple) are probably familiar to most English-speaking American 3- and 4-year-olds, others (e.g., candle and magnet) might not be familiar. If children are unfamiliar with one word, they might perseverate, focusing on the more familiar option or, alternately, adopting a novel label for the atypical object (Merriman & Bowman, 1989).

A complementary verbal factor that might affect AR performance is children's understanding of the predicates ''looks like'' and ''really and truly.'' If even one question is ambiguous, the child might answer the more interpretable question and perseverate on that response for the less interpretable one. Thus, to answer correctly AR questions, children must activate each predicate's intended semantic implications and select the most strongly associated, or implied, label. Success requires inferring the intended predicate meanings as well as related aspects of the relevant word meanings. We tested the effect of lexical knowledge by assessing children's knowledge of labels for AR test objects (Experiment 1) and by examining the association between AR performance and receptive vocabulary (Experiment 2). We tested the role of predicate knowledge by comparing children's performance with standard AR questions with their performance with modified questions that used less ambiguous, and perhaps more semantically distinct, predicates (Experiment 2). We predicted that children with more knowledge of AR object labels (and knowledge of words in general) would perseverate less in the AR test. We also predicted that children would perseverate less if the test questions were more distinct and unambiguous.

In sum, children's tendency to perseverate across questions in certain discourse contexts, and their limited understanding of words and predicates in the AR test, might cause AR errors. These factors are largely independent of object properties such as deceptiveness; therefore, if these hypotheses are

correct, they predict that (a) children's perseverative AR task error rates will be replicated in a control task with nondeceptive objects and different questions, and (b) measures of lexical knowledge (e.g., receptive vocabulary) will account for variance in children's performance. A separate question is whether verbal working memory limits cause representational inertia that leads to perseverative AR errors. If so, children's word memory span should predict their AR errors, and object deceptiveness should moderate children's AR error rate or error type.

Experiment 1

Several questions about sources of variance in preschool children's object AR errors were addressed. First, which verbal or representational abilities (if any) predict errors? Two abilities were studied. One is knowledge of words for AR objects. Perhaps uncertainty about the word for either the appearance or function of a deceptive object facilitates perseveration (e.g., to avoid an unfamiliar word). The second is verbal working memory span. Memory limits might cause AR errors because choosing each word correctly requires maintaining, for several seconds, accurate representations of the most recent question and word options. These memory demands seem to fall near the limits of 3- and 4-year-olds' working memory span, as measured by laboratory tests (Gathercole & Adams, 1993; Gathercole, Service, Hitch, Adams, & Martin, 1999).

Second, do AR errors stem from a general tendency to perseverate when answering successive questions about an object or array? To answer this question we administered a control test with nondeceptive items, as well as the standard AR test. Both tests pose two successive questions with distinct predicates, each of which establishes a choice between two valid labels. Both tests therefore entail selective predicate ⇔ label mapping. Control test stimuli were pictures of animals holding or wearing discrete objects (e.g., a duck wearing a hat). Children were asked what the animal ''looks like'' and what it ''has.'' A general perseverative tendency would produce similar AR and control test error rates, independent of variance due to age and verbal working memory.

Third, does object deceptiveness influence the form or frequency of perseveration? If children are subject to representational inertia, they should be sensitive to deceptiveness because perceptual cues from highly deceptive objects should activate one kind of label (i.e., appearance words) more than the other (i.e., function words). This might be expected to facilitate more phenomenism errors. ''Bad fakes''—less-deceptive objects—might have salient function cues that create more competition with the appearance term, thereby facilitating more realism errors. Alternately, high-deceptive objects might elicit more perseveration overall because asymmetry in the availability of perceptual cues to each category (or label) makes it hard for children to keep both labels equally active. In any case, if representational inertia makes the AR test difficult for preschool children, the form or frequency of their perseveration should be moderated by object deceptiveness. To test this possibility, every child completed the AR task with both low- and high-deceptive objects.

Method

Participants

Twelve 3-year-olds (5 girls; age: $M = 3$ years, 7 months; *range* = 3,0–3,11), twelve 4-year-olds (5 girls; age; $M = 4,6$; *range* = 4,1–4,11), and twelve 5-year-olds (3 girls; age: $M = 5,6$; *range* = 5,1–5,9) were tested. In addition, 10 adult college students (6 women) and twelve 6- and 7-year olds (9 girls; age: $M = 6,9$; *range* = 6,2–7,5) participated in the object deceptiveness pretest. The mostly White, middle-class participants were recruited from Vanderbilt University classes and from nearby Nashville, Tennessee, preschool and after-school programs.

Materials

A set of 27 candidate objects (details available from the authors) were rated on deceptiveness in a pretest (procedure and results as described later). The results yielded three high-deceptive items and three low-deceptive items with matched functions, allowing control over participants' familiarity with object functions, and function labels, across levels of deceptiveness. The high-deceptive objects were a candy magnet, peanut eraser, and lipstick pen. The low-deceptive objects were a banana magnet, strawberry eraser, and carrot pen.

The word knowledge test used real exemplars of each appearance and function category from AR test objects (e.g., a real strawberry, a ball-point pen). Exemplars were selected for prototypicality and ease of identification, as judged by the researchers. The control test used six laminated photographs of animals holding or wearing distinctive items: a duck wearing a hat, a horse with keys in its mouth, a

monkey holding cookies, a bear holding a ball, a bunny with an apple, and a cat with grapes.

Deceptiveness Pretest Procedure and Results

After several practice items, children were asked to identify what each candidate object really was and then place it into one of three boxes, labeled *very tricky, sort of tricky,* and *not very tricky.* Adults, after several practice items, rated each candidate object on a 7-point Likert scale, ranging from 1 (*really good fake*) to 7 (*really bad fake;* procedural details are available by request from the first author). Both child and adult participants were allowed to examine each candidate object (presented in random order) before sorting or rating it. All participants in both age groups seemed to understand the tasks. The two age groups agreed about which objects were more or less deceptive (between-age correlation between object ratings: $r = 0.82$). Children rated the three high-deceptive objects as significantly trickier than the low-deceptive objects, $t(11) = 5.0$, $p = .001$, and rated both sets as different from the overall mean, low set: $t(11) = -4.1$; high set: $t(11) = 4.8$, both ps = .003. Adults also rated high- and low-deceptive subsets as different from one another, $t(9) = -12.4$, $p = .001$, and from the mean of all objects, $t(9) = -13.2$ and $t(9) = 7.2$, both ps = .001.

Procedure

All four tests were administered in a single session in a fixed order: (a) AR, (b) memory span, (c) control, and (d) word knowledge.

AR test. As in Flavell et al. (1986), children saw a deceptive object and labeled it (by appearance). Its function was then demonstrated and named. Children then answered an appearance and a reality question: ''Does it look like a peanut or does it look like an eraser?'' and ''Is this really and truly a peanut or really and truly an eraser?'' Question and word order were counterbalanced. Object order was randomized with the constraint that two same-function items were not presented consecutively.

Memory span test. Children heard two lists of six words (alphabet, arm, driver, hate, holiday, letter; and bird, button, newspaper, picture, pot, potato) and were encouraged to recall as many as possible. List and word order were randomized. To make the memory demands similar to the AR test, chosen nouns had similar length and frequency as AR object labels. Children heard and recalled each list twice.

Control test. Children saw photos of animals holding or wearing familiar items (e.g., monkey with a cookie). The child named the animal (e.g., ''What does this look like?'') and its possession (e.g., ''What does it have?'') and got feedback. The child was then asked, for example, ''Does it look like a monkey or does it look like cookies?'' and ''Does it have a monkey or does it have cookies?'' Predicate and word order were counterbalanced, and picture order was randomized.

Word knowledge test. Children saw real exemplars of the appearance and function category of each AR object (e.g., real peanut and eraser). They were prompted to name each exemplar (''What is the name for this?'' ''What do you call this?''). If they did not answer within 10 s, two more prompts were given: First, ''It's not a [foil word], right? What is it?'' If this did not elicit a response, a forced choice was requested: ''Is it [foil word 2], [correct word], or [foil word 3]?''

Results

Preliminary analyses of gender differences in each dependent variable revealed no differences that approached conventional levels of statistical significance, with one exception: Girls performed significantly better than boys in the labeling test (Ms = 7.8 vs. 7.0, SDs = 1.3 and 1.4, respectively), $F(1, 33) = 4.7$, $p = .038$. This is consistent with findings that when gender differences in young children's language skills occur, they favor girls (e.g., Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). Because there were no meaningful gender effects in any other test, however, we did not enter the factor in any further analyses.

AR Test: Age and Object Deceptiveness

The number of objects (zero to six) to which children responded correctly (to both questions) was entered in a mixed-measures ANOVA, with age (3, 4, or 5 years old) as between-children and deceptiveness (high and low) as within-children factors. Age had a significant effect (see Table 1), $F(2, 33) = 4.6$, $p = .017$. Post hoc Sheffé tests showed that 5-year-olds made significantly fewer AR errors than did 4-year-olds ($p = .036$). However, 3-year-olds did not make significantly more errors than either 4-year-olds or 5-year-olds, though the latter difference approached significance ($p = .055$). It is curious that 4-year-olds did not outperform 3-year-olds, though this might reflect sampling error: Four-year-olds at the preschool sites where we recruited greatly outnumbered 3-year-olds; perhaps some parental selection factor (e.g., informal assessment of

Table 1
Mean Numbers (and Standard Deviations) of Correct Responses to Appearance–Reality (AR), Control, Word Knowledge, and Memory Span Tasks, by Age: Experiment 1

	Age group		
Task	3	4	5
AR	1.9 (1.8)	1.7 (2.2)	4.1 (2.2)
Control (predicate matching)	1.7 (1.9)	2.6 (2.1)	3.7 (2.2)
Word knowledge	6.5 (2.2)	7.4 (1.1)	8.1 (1.0)
Memory span	0.9 (0.7)	1.7 (0.6)	1.6 (1.2)

Note. AR and control tasks: means out of 6 correct patterns; word knowledge: means out of 12 AR labels; memory span: words recalled out of 6, averaged across four trials.

cognitive and social elements of the child's school readiness) disproportionately affected the 3-year-old sample. This speculation raises intriguing questions for future research, but for now the issue is not critical because our research questions did not hinge on finding significant AR performance differences between 3- and 4-year-olds.

The effect of deceptiveness was not significant, $F(1, 33) = 3.1$, $p = .088$, and the means differed in the opposite direction than was predicted by a representational inertia account. Children produced means (and standard deviations) of 1.4 (1.3) and 1.2 (1.2) out of 3 correct responses to high- and low-deceptive objects, respectively. The Deceptiveness × Age interaction only approached conventional levels of significance, $F(2, 33) = 3.2$, $p = .054$.

Phenomenism errors decreased with age, consistent with previous reports: Means were 2.1 ($SD = 1.7$), 1.7 ($SD = 2.2$), and 0.1 ($SD = 0.3$) in 3-, 4-, and 5-year-olds, respectively, $F(2, 33) = 5.1$, $p = .01$. Post hoc Sheffé tests showed that 5-year-olds made significantly fewer phenomenism errors than 3-year-olds, $p = .018$, but not significantly fewer than 4-year-olds ($p = .071$); 3- and 4-year-olds did not differ. Realism error rate did not change significantly with age: Means were 1.1 ($SD = 1.2$), 1.8 ($SD = 2.2$), and 1.7 ($SD = 1.9$) in 3-, 4-, and 5-year-olds, respectively, $F(2, 33) < 1$. Paired t tests showed no significant differences in phenomenism or realism errors with either high- or low-deceptive objects. This does not support the idea that deceptiveness of objects is central to perseverative AR errors.

Memory Span Test

Word spans (averaged over four trials) increased from 0.9 in 3-year-olds to 1.7 in 4-year-olds and 1.6 in 5-year-olds, $F(2, 33) = 3.3$, $p = .048$. None of the group differences achieved conventional significance levels in post hoc Sheffé tests. Children recalled more words after hearing the list a second time, suggesting they were attentive and motivated.

Control Test

The increase in mean correct responses (*range* = 0–6) with age (Table 1) was not reliable, $F(2, 33) = 2.6$, $p = .091$. Children perseverated on means (and standard deviations) of 1.2 (1.7) animal words and 1.3 (1.6) part words. Neither mean changed significantly with age.

Word Knowledge Test

Children received 1 point for each correct term produced in response to the first two prompts and .5 point for each correct choice from a final forced-choice prompt. As shown in Table 1, word knowledge means increased from 6.5 (out of 12) in 3-year-olds to 8.1 in 5-year-olds. This trend approached conventional levels of significance, $F(2, 33) = 3.3$, $p = .051$.

Relations Among Tests

Correlations among number of correct response patterns in the AR test and correct patterns in the control test, word knowledge test, and memory span test, with age partialled out, are shown in Table 2. The partial correlation between correct AR and control test responses, $r(32) = .58$, is significant, $p = .001$, as are partial correlations between correct AR responses and word knowledge, $r(32) = .41$, $p = .013$, and the control test and memory span, $r(32) = .38$, $p = .025$. The partial correlation between number of perseverative responses in the AR and control tests, with age, word knowledge, and memory span controlled, remains significant, $r(31) = .60$, $p = .001$. A forward stepwise regression confirmed that 53% of AR variance was accounted for by control test perseveration, $R^2 = .38$, $F(1, 34) = 21.1$, $p = .001$, plus word knowledge, R^2 change = .15, $F(1, 33) = 10.4$, $p = .004$. No other factor explained additional significant unique variance. Because memory span scores are skewed, a nonparametric test of their relation to AR performance was conducted by classifying children as high, medium, and low performers on each test. The distribution did not significantly differ from the expected distribution, $\chi^2(4, N = 36) = 5.5$, $p = .24$.

Table 2
Partial Correlations Among Correct Appearance–Reality (AR) and Control Test Responses, Word Knowledge, and Verbal Memory Span, With Age Controlled: Experiment 1

	Task		
Task	Control	Word knowledge	Memory span
AR	.58**	.41*	.11
Control		.32	.38*
Word knowledge			.24

Note. Word knowledge: labeling test number correct; verbal memory span: mean words recalled.
*$p < .05$. **$p < .001$.

An obvious question is whether children's word knowledge predicts specific AR errors. If a child failed to name an exemplar (e.g., peanut) in the word knowledge test, did he or she then perseverate on the other word (i.e., eraser)? To address this question, we coded each AR response for whether the correct word was chosen. These 12 responses (6 objects × 2 words) were cross-classified with the child's response to the corresponding exemplar in the word knowledge test (saying the name by the second probe was coded as correct). Most AR errors were omissions of words that were produced in the word knowledge test, $t(35) = 2.6$, $p = .02$ (two-tailed). The analysis does not adjust for base rates, however, which is relevant because children correctly named most exemplars. Thus, although there is no compelling evidence that failure to produce specific words predicts specific AR errors, we cannot reject the possibility.

Discussion

Most accounts of object AR errors have focused on children's ability to grasp changes in mental content (i.e., beliefs), or to represent dual identities. These accounts lack compelling empirical support. In contrast, the current study shows that children's object AR performance is predicted by their tendency to perseverate in a seemingly unrelated task (e.g., reporting what an animal "looks like" vs. what it "has") and their knowledge of object labels.

These findings complement claims that linguistic factors (Sapp et al., 2000), and more specifically pragmatic factors (Rice et al., 1997; Siegal, 1991), account for developmental changes in AR performance. Specifically, the tendency of some children to perseverate in labeling objects, in both the AR and control tests, suggests that blending or "leaking" responses across questions can affect a wide range of verbal tasks. Word knowledge seems to mediate this effect; perhaps perseveration is triggered by a child's unfamiliarity with the meanings of the words he or she is asked to choose between.

One caveat to this account is that both the AR and control tests used the predicate "looks like a …" Perhaps the correlation between tests stems from individual children's comprehension of this predicate, rather than a general perseverative tendency. To test the former explanation, we calculated the correlation between perseverative responses to "looks like" questions in both tests, and the correlation between perseverative responses to the other question ("really a …" or "has a …") in both tests. Age, memory span, and word knowledge were partialled out. The association between tasks in "looks like" perseveration, $r = .30$, is neither significant ($p = .091$) nor stronger than the association between perseveration on dissimilar questions, $r = .40$. Thus, there is no evidence that the common predicate accounts for the between-test correlation. Nevertheless, we address the possibility further in Experiment 2 by adding a second control task with no overlap in predicates and by experimentally manipulating the AR questions.

Word memory span did not predict AR performance. This is noteworthy because age-related limits in working memory could explain why presenting visible exemplars (e.g., typical peanut) improves children's AR performance (Rice et al., 1997). Before concluding that there is no relation between individual children's AR performance and verbal working memory capacity, however, we must thoroughly test the relation. If the two are indeed independent, we need another explanation of why visible exemplars help children in the AR task. Perhaps, for example, exemplars serve not as memory cues but as cues to the presence of two response options for both test questions. That is, making the word choices concrete (i.e., embodied as exemplars) highlights the indeterminacy of the questions—the fact that each question is a separate problem. If children encode each question as independently indeterminate, so that the answer to one does not automatically extend to the other, they should be less likely to perseverate. This possibility is explored in Experiment 2.

What do these data tell us about verbal abilities and AR performance? On the one hand, the positive effect of understanding specific object labels might reveal a critical role of semantic mapping (i.e., choosing labels based on the implications of the predicate of each question). On the other hand, the relation between word knowledge and AR errors

was only correlational, not tied to lack of knowledge of certain words; therefore, general lexical and semantic knowledge (including processes of choosing between nuanced semantic associations) may predict children's performance. Consistent with this, Deák and Maratsos (1998) and Deák et al. (2001) found that receptive vocabulary predicts 3- to 5-year-olds' productivity in labeling the appearance and function of representational objects. To test whether general lexical-semantic knowledge predicts flexible word selection in the AR test, Experiment 2 included a receptive vocabulary test.

A related question is whether AR errors stem from failure to comprehend an AR question. For example, the word ''truly'' in the predicate ''really and truly'' probably does not imply object identity to young children. Also, preschoolers seldom use ''real'' or ''really'' to contrast reality with illusion (Woolley & Wellman, 1990). Finally, many 3- and 4-year-olds do not understand the predicate ''looks like a…'' (Deák, 2000). These findings suggest that some AR questions are hard for preschoolers to understand. In Experiment 2 we tested whether specific predicates facilitate children's AR errors by comparing performance under standard and modified questions.

Finally, in Experiment 1, object deceptiveness did not have a significant impact on AR performance. Although we might have lacked adequate power to detect a deceptivenes effect, the only trend—toward more errors on less deceptive objects—is not predicted by a working memory account or by any obvious consequence of representational inertia. We do not believe the negative evidence was due to a weak manipulation: Our bad fakes were obvious, even from a distance, whereas the good fakes continued to fool some adults even while handling them. J. Heberle (personal communication, February 2002) also found no effect of object deceptiveness on AR errors, independently confirming our finding. The simplest conclusion is that AR errors are not reliably related to the deceptiveness of AR stimuli—the relative perceptual availability of cues to a representational object's function has little to do with children's naming errors.

Experiment 2

The results of Experiment 1 leave several unresolved questions about the relation between children's AR responses, verbal abilities, and perseveration. First, the correlation between the AR and control tests suggests a general perseverative trait. Perseveration might be elicited by successive questions about a complex object, each of which demands choosing between valid labels or descriptions. It is possible, however, that specific verbal content is critical for the correlation because both tests used the predicate ''looks like a….'' Perhaps children perseverate because of confusion about this predicate's meaning.

To determine whether ''looks like a…'' questions contribute to perseverative errors, two new control tests were designed. In the overlap control test children answered two questions about nondeceptive test objects. One question used the predicate ''looks like a….'' In the nonoverlap control task, children answered two questions about different nondeceptive objects, but neither question used the predicate ''looks like a….'' If the correlation in Experiment 1 was due to the common predicate rather than to a general perseverative trait, children's performance on the overlap control and AR tests should be more strongly correlated than their performance on the nonoverlap control and AR tests.

The second question concerns the specific types of verbal skills or knowledge that contribute to AR performance. One idea is that predicate mapping is critical for accurate, flexible label selection. In the AR test children must selectively associate or map each predicate (''looks like'' or ''is really and truly'') to a candidate label (e.g., peanut or eraser) with regard to a specific object. Perseveration might result from mapping nonspecificity: either not knowing the implications of each predicate or not knowing which stimulus property warrants each label. In either case, the effect would be perseverative responses, which can be defined as failure to map specific predicates to specific labels. The putative shift from less to more specific predicate mapping is depicted in Figure 1: Younger, less verbal children tend to perseverate in any test that requires using specific semantic implications to select labels. Such tests include, of course, both control tests and the AR test.

The mapping specificity account implies that lexical knowledge mediates AR performance. Children's ability to draw the denotative and connotative implications of words and phrases (rather than, say, syntactic knowledge) should predict mapping specificity. Vocabulary is a convenient index of this ability; a receptive vocabulary test that predicts overall age-normed verbal inference skill is the Peabody Picture Vocabulary Test–Revised (PPVT–R; Dunn & Dunn, 1981; Sattler, 1988). We tested whether PPVT–R scores predict AR performance. An alternative hypothesis, that knowledge of specific AR object labels is critical, would imply that the

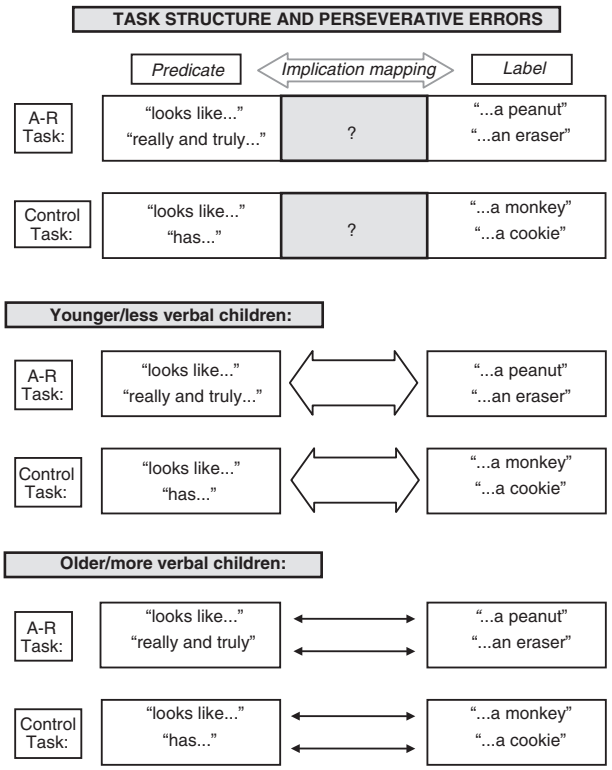

Figure 1. Schematic representation of the hypothetical shift from less specific to more specific predicate↔label mapping, as a result of growing comprehension of predicate implication, and of attributes denoted by each label, with respect to a stimulus.

PPVT is too blunt an instrument to predict significant variance in children's AR errors.

Another alternative hypothesis is that comprehension of specific predicates "really and truly a…" and "looks like a…" determines children's AR performance. To test this hypothesis, standard AR questions were compared with modified questions with different predicates. The alternate questions "Do you *really use it as* a…?" and "Is its *shape and color like* a…?" were chosen because they more specifically denote the dissociation in deceptive objects. Function and appearance typically are redundant in artifacts, and although they are dissociated in deceptive objects, thorough dissociation is virtually impossible. That is, perceptual data are our principle means for identifying an entity. When initial perceptual evidence is ambiguous, we enhance it by exploration: leaning over a bubbling pot and breathing deeply to identify an aromatic spice, peering intently at a distant tree to identify a perched bird—even using "perceptual prosthetics" (e.g., jeweler's eyepiece) to gain acuity. During these or any exploratory processes, our perceptions change (Gibson & Gibson, 1955), and as a consequence, so do the probabilities of assigning a percept various categories and labels. Thus, a candle shaped and colored like an apple looks like an apple only under superficial inspection. It actually looks only roughly like an apple; its size, shape, and texture are all a bit wrong. Similarly, it is really and truly an artifact with two functions: to give light and to look like an apple. Because appearance and reality are only partly dissociated, standard AR questions require nonliteral interpretation. Young children's ability to make nonliteral interpretations is limited (Campbell & Bowe, 1983; Olson, 1977). We designed alternative AR questions to capture more accurately the partial dissociation embodied in deceptive objects: between global appearance and primary intended function. We expected our alternate questions to be less ambiguous than the standard questions and therefore to enhance mapping specificity and reduce perseveration.

A lingering question is whether working memory predicts variance in AR performance. Though data from Experiment 1 answered this question in the negative, memory span scores in that sample were low and skewed, and some incidental memory demands were not controlled between the AR and word span test. Thus, children completed a new word memory test designed to match the memory demands of the AR test more precisely and to elicit a wide range of recall capacity.

Experiment 2 also evaluated the hypothesis that failure to notice indeterminacy accounts for AR errors. Detecting indeterminacy means realizing that the current answer or response is not determined or given by prior responses. Realizing this is crucial for solving many problems and making many inference. Children younger than 6 years, however, often treat ambiguous messages, questions, or tasks as clear and determinate (e.g., Markman, 1979), indicating a general insensitivity to indeterminacy. However, sensitivity to indeterminacy varies across preschool children (e.g., Patterson, Cosgrove, & O'Brien, 1980), and this variability might contribute to performance in the AR test, where children must notice that each of the two questions per object poses a unique indeterminacy problem (i.e., of meaning selection). The answer to one question does not determine the next; each predicate's implication must be drawn separately for a predicate↔word mapping. If questions are perceived as redundant, or interdependent rather than independent, children might assume they know the correct response to all questions after the first and perseverate as a result. Thus, insensitivity to indeterminacy might explain Rice et al.'s (1997) finding that exemplars improve 3-year-olds'

AR responses. That is, embodying word choices during the AR test might highlight the demand to make a choice for each question. By this logic, if exemplars were presented but obscured right before the experimenter asked the questions, the exemplars should highlight the indeterminacy of each question. However, if exemplars facilitate AR performance by providing memory cues (Rice et al., 1997), they should help only if they remain visible while children respond. We tested these alternatives by putting exemplars into distinct, opaque boxes before asking children the AR questions. In this situation, exemplars do not serve as available cues to a specific function or appearance label for the test object, but they highlight the need to choose between responses for each question. This manipulation helps to narrow down the cause of exemplar facilitation. Performance in the hidden exemplar condition was compared with a control condition in which the two boxes, without exemplars, simply sat on the table during the test.

As a more direct test of children's sensitivity to indeterminacy, and its relation to AR performance, we designed a detection of indeterminacy test. Children saw several scenarios with variable outcomes (e.g., location of a penny, color of the next poker chip drawn from a box). Each scenario had one version with an obvious, determinate outcome (e.g., penny is hidden in child's view, all chips of the same color), and another version with an uncertain outcome (e.g., penny is hidden before child arrives, chips of many colors). Children judged whether the outcome of each scenario is determinate or indeterminate. If AR errors stem from insensitivity to indeterminacy, performance on this test and the AR test should be positively associated.

Method

Participants

Sixty-four preschoolers participated: thirty-two 3-year-olds (15 girls; age: $M = 3,6$; *range* = 3,1–3,11) and thirty-two 4-year-olds (14 girls; age: $M = 4,7$; *range* = 4,0–4,11). Children were recruited from the same population as Experiment 1.

Materials

The training item for the AR test was a car-shaped crayon. Test items were an apple-shaped candle, a shell-shaped soap, a lipstick-shaped pen, a book-shaped box, a peanut-shaped eraser, and a candy-shaped magnet. These items were intermediate or high in deceptiveness (as rated in Experiment 1). Real, prototypical exemplars of each appearance and function category (e.g., red apple, candle) were used, as were two different-colored, opaque boxes.

Two sets of six objects were used for the two control tests. Each object had a familiar shape, material, and part (e.g., a wooden house shape with a small bow attached). Shapes were teddy bear, duck, flower, heart, house, and star. Materials were fur (imitation), glass, metal, paper, clay, and wood. Parts were a bell, a bow, a button, a flower, a key, and a zipper.

The detection of indeterminacy scenarios used a penny; a transparent bingo ball (i.e., 20 cm plastic sphere mounted on a pivot and spun by a handle, so that one small item falls from the ball per rotation); marbles of many colors; two spinners for a children's game (one with pictures of Winnie the Pooh all around it, the other with pictures of different A. A. Milne characters); and two ''presents,'' one wrapped and the other unwrapped and open.

Procedure

Each child participated in two sessions, no more than 1 week apart. Three tests were given in each session. Test order was quasi-random: The two control tests and the AR and detection of indeterminacy tests were always given in different sessions. To hold children's interest and provide a break from testing, children played with stickers and a coloring book between tests.

AR test. Children first answered AR questions about a training object, with feedback. They then saw six test objects in random order. Half of each age group answered standard AR questions (i.e., ''What does this look like?'' and ''What is this really and truly?''). The other half answered modified questions (i.e., ''Do you really use it as …'' and ''Is its color and shape like ...;'' see the Appendix). Half of each question group saw appearance and function category exemplars (e.g., real apple, candle) while learning about the deceptive object, as in Rice et al. (1997). Before the AR test questions were posed, however, exemplars were placed in distinct boxes. As the experimenter listed the two word choices for each question, she pointed to the boxes to indicate the connection between the questions and the hidden referents. In the control condition, the same boxes were on the table during the test but were never used, and no exemplars were shown. Question and word order were counterbalanced within each condition.

Control tests. Each child completed both control tests. In each, children first discussed aspects of the object (analogous to the AR test), then answered two questions (see the Appendix). In both tests, one question per object used the predicate ''have a…'' regarding a distinctive part. In the overlap control test, the other question used the predicate ''looks like a…'' regarding the object's shape. In the nonoverlap control test, the other question used the predicate ''is made of…'' regarding the object's material. Note that 3- and 4-year-olds are above chance in mapping novel words predicated by ''is made of…'' onto novel materials, and novel words predicated by ''has a…'' onto novel parts (Deák, 2000). Object sets were randomly assigned to tests, object order was randomized, and word and question orders were counterbalanced.

Indeterminacy detection test. Children were given four problem scenarios, each with two versions: indeterminate and determinate. Children were asked to judge, for each version, whether they ''know for sure'' or ''have to guess'' the answer or outcome. (Preschool children discriminate the verbs *know* and *guess* with regard to implied certainty; Moore, Bryant, & Furrow, 1989). If children did not readily respond ''know'' or ''guess'' or the equivalent (e.g., if the child tried to predict the outcome), the experimenter restated the question. The scenarios concerned: (a) the number of fingers the experimenter was holding up, either visibly (determinate) or out of sight (indeterminate); (b) the color of the next marble to drop from a hollow, rotating sphere (i.e., bingo ball) filled with either marbles of one color (determinate) or marbles of many colors (indeterminate); (c) which cartoon character a spinner's arrow would point to when it stops, given either seven pictures of the same cartoon character (determinate spinner) or pictures of seven different characters (indeterminate spinner); and (d) the contents of a gift box that was either unwrapped and open (determinate) or closed, wrapped, and tied with ribbon (indeterminate).

Children were trained first on the task format and questions. The training situation concerned the location of two pennies: one visible to the child (determinate) and the other described as hidden somewhere ''… in this *big* building!'' (indeterminate). Children received feedback and explanation on the training scenario and then completed four test scenario pairs (without feedback) given in random order. Order of probe choices (''know for sure'' or ''have to guess''), and scenario version (determinate or indeterminate) was counterbalanced.

Vocabulary and memory span test. The PPVT–R was administered using the standard procedure (see Dunn & Dunn, 1981). Verbal memory span was tested in an immediate list recall paradigm, using four lists, each with four different words (potato, letter, vine, Sunday; newspaper, driver, arm, pencil; telephone, picture, pot, flower; holiday, bird, button, dishes). Words roughly matched AR object labels for length and frequency. After a practice trial, each list was read twice (to equate for word repetition in the AR test and to reduce skew of recall scores), and children were immediately encouraged to say back as many words as they could remember. Children received stickers between lists, both to match the AR intertrial interval and to reduce proactive interference between lists. List and word order were randomized.

Results

Overall Age and Gender Patterns

Preliminary analyses of gender differences revealed no significant effects; therefore, data from boys and girls are combined in subsequent analyses. Means in each task from 3- and 4-year-olds are shown in Table 3, with significant age differences ($p < .05$, two-tailed t tests) indicated. In contrast to Experiment 1, there is robust, predictable improvement from 3 to 4 years. Also, PPVT–R means and variance are not statistically different from published norms (Dunn & Dunn, 1981), suggesting our sample has typical verbal skills.

AR Test

Mean frequency of each response pattern (correct, phenomenism, realism, switched) are shown in Figure 2 for each age group and exemplar condition. Phenomenism declined with age, and realism was less frequent in the no-exemplar control condition. A MANOVA comparing frequency of each perseverative response type (within participants) with age and exemplar condition (between participants) revealed a significant multivariate age effect, $F(2, 59) = 5.5$, $p = .007$. This effect was due to a significant reduction in phenomenism with age, $F(1, 60) = 9.7$, $p = .003$. A reliable multivariate exemplar effect, $F(2, 59) = 4.5$, $p = .015$, was due to a significant increase in realism errors in the hidden exemplars condition, $F(1, 60) = 7.5$, $p = .008$. A multivariate Age × Exemplar interaction, $F(2, 59) = 5.4$, $p = .007$, is due to fewer phenomenism errors by 3-year-old in the control (i.e., no exemplars) condition and relatively

Table 3
Mean Numbers (and Standard Deviations) of Scores on Tasks by 3- and 4-Year-Olds in Experiment 2

	Age group	
Task	3	4
Appearance–reality (AR)*	1.6 (1.7)	3.4 (2.3)
Control: Overlap (''looks like …'')**	3.2 (1.7)	4.7 (2.1)
Control: Nonoverlap (''is made of …'')**	2.5 (1.8)	4.7 (1.9)
Detection of indeterminacy**	1.1 (1.2)	2.3 (1.2)
PPVT–R (age standardized)	103.8 (20.5)	109.2 (13.9)
Memory span**	1.1 (1.0)	2.5 (0.9)

Note. AR and control tasks maximum correct = 6.0; detection of indeterminacy and memory span maximum correct = 4.0; Peabody Picture Vocabulary Test–Revised (PPVT–R) population mean = 100 (SD = 15).
*$p < .05$, two-tailed t test of age differences. **$p < .01$, two-tailed t test of age differences.

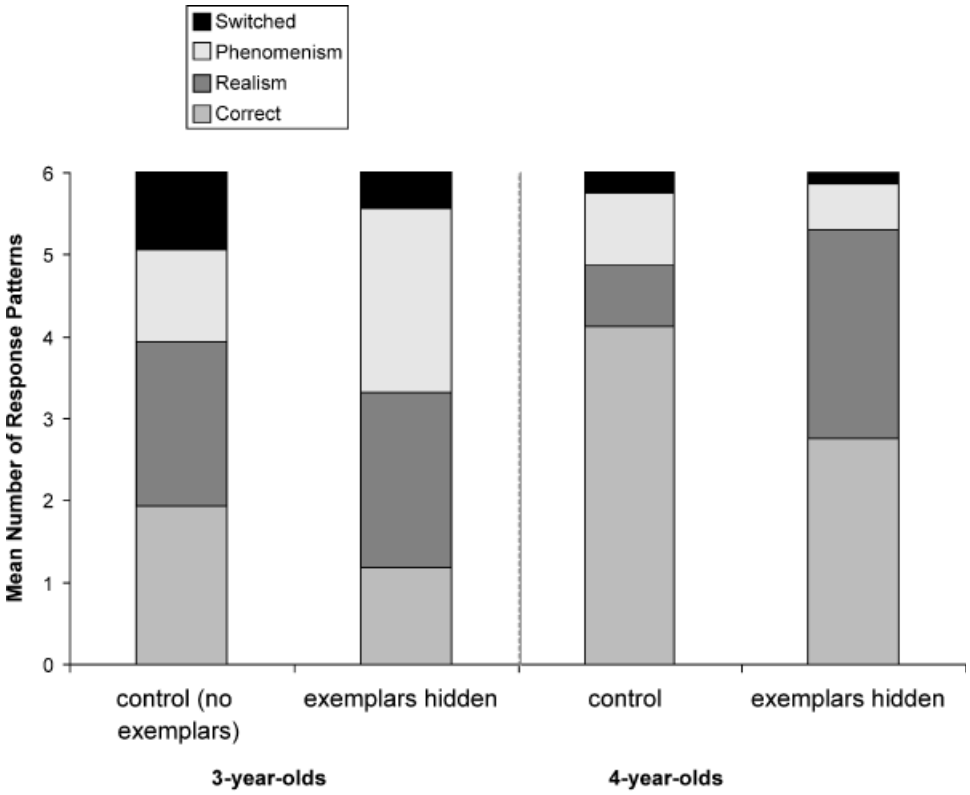

Figure 2. Mean number of correct appearance–reality answer pairs, phenomenism errors, realism errors, and switched (both incorrect) answers, by age and exemplar condition, Experiment 2.

more realism errors by 4-year-olds in the hidden exemplars condition. It seems that presenting exemplars, then hiding them while asking children the AR questions, highlighted objects' appearance for 3-year-olds and highlighted objects' function for 4-year-olds.

Control Tests

Performance on both control tests varied considerably across children (see Table 3), permitting measures of association with other tests. Errors were mostly perseverative, and total perseveration in both

tests declined with age. Errors in the overlap control test (''looks like a …'' and ''has a …'') were almost evenly divided between persisting on the shape word and persisting on the part word. This is notable because preschool children reliably associate ''has a …'' with a specific object property (i.e., parts), but they do not reliably associate ''looks like a …'' with shape (Deák, 2000). Because the former predicate has more specific implications than the latter, we might have found more perseveration on the better specified (i.e., part word) response. Apparently, however, perseverative errors are not predictable from children's understanding of specific predicates.

Further evidence that perseverative errors do not depend on the nonspecific ''looks like a'' predicate comes from the nonoverlap control test, on which children performed no better although the ''looks like'' predicate was not used. Note that 3- and 4-year-olds can interpret the replacement predicate ''is made of'' as referring to material (Deák, 2000); therefore, we did not merely replace one ambiguous predicate with another. In sum, the control tests do not indicate that specific predicates mediate perseveration.

Because control objects were in no way deceptive, the findings add weight to our argument (based on Experiment 1) that perceptual accessibility of cues to both labeled categories does not significantly mediate flexible word selection in a forced-choice, two-question paradigm.

Indeterminacy Detection Test

Children correctly responded (i.e., judged the need to guess indeterminate scenario outcomes and the ability to know determinate scenario outcomes) to a mean (standard deviation) of 1.7 (1.3) out of four situation pairs. They perseverated on 1.1 (1.1) ''guess'' and 1.0 (1.0) ''know for sure'' responses. Thus, children did not consistently overinterpret the determinacy of ambiguous scenarios. Given prior evidence that preschoolers are overconfident in interpreting ambiguous messages or scenarios (e.g., Markman, 1979), the lack of bias toward ''know for sure'' perseveration is surprising. On the chance that some scenarios strongly elicited guess responses, thus counteracting a know bias, we examined children's responses to individual scenarios. One stood out: The determinate spinner outcome (i.e., all stickers of one character) was construed by many children as indeterminate. Perhaps children perceived discrete stickers of the character as representing different individuals, or perhaps they thought the experimenter was asking which sticker the spinner would point to, rather than which individual. Either construal would make a guess response sensible.

Verbal Memory Span Test

Word memory span scores covered the entire range (0 to 4), and 78% of children's mean scores fell between .5 and 3.5 words. Thus, recall varied widely across children, allowing meaningful tests of association.

Relations Among Tests

To understand the relation between performance on the AR task in different conditions and performance on the other tasks, a backward stepwise linear regression analysis of correct AR responses (*range* = 0–6) was conducted. The following variables were entered: age (by 6-month group), AR questions (standard or modified), exemplars (hidden or none/control), working memory span (mean number of words recalled), vocabulary (PPVT–R raw score), correct overlap control responses (*range* = 0–6), correct nonoverlap control responses (*range* = 0–6), and correct indeterminacy detection scenarios (*range* = 0–4). Four variables accounted for half of the variance ($R^2 = .50$): overlap control performance, vocabulary, indeterminacy detection, and exemplar condition. Three of these were significant in the final model: overlap control correct ($\beta = .33$, $p = .005$), vocabulary ($\beta = .28$, $p = .014$), and indeterminacy detection ($\beta = .25$, $p = .024$). All are positively related. Exemplar condition was negatively but not significantly related ($\beta = -.18$, $p = .069$). Other variables (age, memory span, AR questions, and nonoverlap control task) together accounted for only 3% of unique variance.

Perhaps a general perseverative tendency influenced performance on the AR tests as well as the control and indeterminacy detection tests. With regard to the last of these, both situations in a problem pair used similar props (e.g., spinners) and events, and children chose between the same two responses (''know for sure'' vs. ''have to guess''). The procedure therefore resembles the discourse structure of the other tests. To determine whether the indeterminacy detection test and control tests tap into a common perseverative tendency, we computed partial correlations (with age, vocabulary, and memory controlled) among perseveration rate in all three tests. These ranged from $r = .26$ ($p = .042$) to $r = .43$ ($p = .001$), suggesting an underlying general

perseverative tendency that differs across preschool children and that might predict AR test performance.

To estimate each child's perseverative tendency, all perseverative responses in both control tests and the indeterminacy detection test were summed. This composite perseveration score was entered into a backward regression analysis of children's correct AR responses, along with age, AR question type (standard or modified), exemplar condition (hidden or none), vocabulary, and memory span. This analysis yielded a simpler but almost equally predictive model, including three factors with total $R^2 = .46$: composite perseveration ($\beta = -.41$, $p = .001$), vocabulary ($\beta = .34$, $p = .003$), and exemplar presentation ($\beta = -.20$, $p = .042$). Thus, little predictive power is lost by reducing performance on the control and indeterminacy detection tests to one composite perseveration score. The full model accounts for $R^2 = .49$, again indicating that additional factors account for little unique variance.

Perhaps shared variance between tests is driven by specific predicates, given that the AR and overlap control test share one predicate (''looks like'') and two control tests share another (''has a''). The data do not, however, support this possibility: The partial correlation between number of perseverative responses to ''looks like'' questions in the standard AR and overlap control tests was $r = .14$; the partial correlation between the alternate perseverative responses (i.e., nonshape label) in the same two tests was $r = -.07$. The partial correlation between ''has a'' (i.e., part label) perseverative responses in the two control tests was $r = .31$ ($p = .017$); the partial correlation between alternate perseverative responses (i.e., shape or material) in the same tests was $r = .18$. These four concordant partial correlations thus average $r_X = .17$; by comparison, the four discordant partial correlations (derived by switching the response measures between tests) average $r_X = .11$. The modest difference between these means does not support the view that perseveration is driven by specific predicates.

For a more robust test of the relation of memory span to AR performance, three levels of recall scores and AR performance were cross-tabulated, as in Experiment 1. Unlike Experiment 1, the distribution was significantly different from expected, $\chi^2(4, N = 64) = 12.5$, $p = .014$: Many children performed at a similar level (i.e., poor, average, or good) on both tests. Note, though, that the χ^2 analysis conflates factors associated with both tests, including age and vocabulary.

Discussion

The Experiment 2 data suggest that young children's performance in the object AR test is associated with general linguistic and question-answering skills rather than the ability to represent changing beliefs about identity or to hold multiple representations of an object. A tendency to perseverate across successive verbal forced-choice questions accounts for the largest portion of unique variance in children's AR performance. Because this tendency is elicited by tests that do not use deceptive stimuli (e.g., the control tests) and questions that are not about appearances, realities, or identities, there is little reason to believe that AR stimuli or questions are central to children's AR errors.

Children's lexical knowledge predicts significant, unique variance in AR performance. Receptive vocabulary, assessed with the PPVT–R, is about as predictive as the more specific word knowledge test in Experiment 1 (though, of course, between-experiment comparisons must be treated cautiously). Notably, receptive vocabulary subsumes all of the variance predicted by age and more. That fact is not surprising if we consider (a) the PPVT–R is a good index of overall verbal intelligence (as well as full-scale IQ); (b) the standard AR test requires verbal interpretation, inference, and selection—all central aspects of verbal intelligence; (c) verbal knowledge varies widely across same-age children (Bates, Dale, & Thal, 1995; Nagy & Herman, 1987); and (d) chronological age is a loose index of cognitive skill, whereas vocabulary is a product (and predictor) of accumulated inferential activity, sociocultural experience, and conceptual growth—all factors that are indirectly important for correct performance on the AR (and control) tests.

The data do not indicate that understanding a specific predicate determines AR performance. This was a reasonable possibility, given that the predicate ''looks like a…'' in both the AR and control tests is ambiguous. It can signify a same-kind judgment, an opinion about resemblance, a metaphoric comparison, or hedging (Lakoff, 1973). Many 3- and 4-year-olds do not interpret novel object labels predicated by ''looks like a…'' in a consistent way, suggesting that they do not know what the predicate implies (Deák, 2000). However, children in Experiment 2 did not consistently perseverate either on responses to ''looks like a…'' or on responses to the other predicate in the question pair. Also, modifying AR questions to be more specific and conceptually accurate did not improve children's AR performance, consistent with Flavell, Green, et al.'s (1987)

claim that phrasing the AR questions does not cause children's errors. Of course, the modified questions might be ambiguous, too, for some reason. (By analogy, consider that differential equations can be more specific and veridical, yet harder to understand, than discrete equations describing the same system.) The best interpretation for now is that AR performance depends on a task set to select independent verbal responses for different questions, but whether a child adopts that task set does not depend on the specific semantic content of the questions.

The latter conclusion needs qualification: Children's ability to interpret AR questions, with respect to candidate labels, probably is related to the predictive value of receptive vocabulary. Certainly at the extremes, question difficulty affects children's ability to select a verbal response. Also, the depth of children's knowledge about the candidate response words will affect their ability to make use of the implications of the questions. Some very young children might not adequately grasp any predicate (standard or modified), and these children probably have limited understanding of the labels as well. The same children, of course, are unlikely to recall many words from a short list and therefore will be among the lowest performers in a recall test, as well as in an AR test.

The data do not indicate that verbal memory span can predict or explain AR performance, though the verbal memory span task in Experiment 2 was matched for AR word familiarity, length, and task demands (e.g., repetitions, between-trial interval). Unlike Experiment 1, though, children's AR and memory span performance levels were significantly related in a nonparametric analysis. This analysis did not, however, account for known confounding factors, particularly vocabulary (as described previously). A reasonable interpretation is that the association between working memory and AR test performance is real but spurious.

An intriguing finding is that showing children response exemplars, then hiding them before asking the AR questions, negatively affected AR performance. In previous studies (Brenneman & Gelman, 1993; Rice et al., 1997), exemplars that remained visible improved performance. It seems, then, exemplars only facilitate performance if they remain visible during the questions; otherwise, they prove distracting, perhaps, or confusing. Perhaps, as Rice et al. (1997) suggest, exemplars reduce working memory demands in the AR test. That does not fit our conclusion that children's memory span and AR performance are not meaningfully related. Of course, it is possible that the exemplar condition reduces some memory demand that is not measured by the word memory span task. An alternative explanation is that visible examples serve as excellent cues to the word choices and therefore encourage or remind children to evaluate each word with respect to each question. By this explanation, the exemplars facilitate selective predicate mapping if they remain perceptually available. Finally, the finding that hidden exemplars impeded AR performance fails to support our hypothesis that hidden exemplars highlight the indeterminacy of the AR questions, thereby discouraging perseveration.

Does ability to detect indeterminacy predict AR performance, in spite of the unexpected exemplar effect (i.e., hiding them impairs children's performance)? Perhaps. Indeterminacy detection test performance made a significant contribution to the full regression model. Children who fail to notice that questions are ambiguous (e.g., Cosgrove & Patterson, 1977; Markman, 1979; Speer, 1984) might automatically tend to produce practiced or prior answers without processing the current question. A more parsimonious account, however, is that indeterminacy detection errors, which were mostly perseverative (i.e., repeatedly choosing "have to guess" or "know for sure" answer for both versions of a situation) manifested a general perseverative tendency. That tendency, when estimated by totaling all perseverative errors across the control and indeterminacy detection tests, contributed significantly to a regression model of AR performance and yielded a more parsimonious model. We therefore infer that indeterminacy detection predicted AR performance simply because both tests evoked a general perseverative tendency. It is still possible that encoding successive questions as different and indeterminate is necessary for flexibly selecting responses, but the current data do not show a strong, specific relation between individual children's AR performance and sensitivity to indeterminacy.

General Discussion

The current findings indicate that children's object AR errors stem in large part from verbal and conversational factors. Children's tendency to perseverate in choosing a word or phrase to answer successive questions, plus limited lexical knowledge, account for as much as half of the variance in 3- to 5-year-old children's AR performance. These findings confirm Sapp et al.'s (2000) finding that 3-year-olds' AR errors are a consequence of the verbal choice task.

A historical explanation for AR errors is that representational inertia prevents young children from flexibly shifting between, or maintaining active representations of, two or more concepts for an entity. Although evidence shows that preschool children can sometimes flexibly shift their representations of an entity (e.g., Deák & Maratsos, 1998; DeLoache, 1995), it is plausible that working memory limitations cause a limited form of representational inertia, which restricts the number of concept labels available (e.g., labeling) within a brief period. We found little compelling evidence, however, that verbal memory span predicted AR performance (though the question remains why hidden exemplars did not facilitate children's AR performance). However, a reader pointed out a potential confound: Several words in both working memory test lists were abstract nouns or event labels rather than concrete nouns. Perhaps this difference ameliorated the association with the AR test, which involves concrete nouns. This possibility could be tested in future studies.

A plausible representational inertia account implies that object deceptiveness should mediate children's difficulty in representing multiple labels or concepts, and thus AR errors. That is, the relative salience of perceptual features that correspond to one or the other concept should influence which label is retrieved first and is repeated. Thus, highly deceptive objects should elicit more phenomenism errors, and less deceptive objects should elicit relatively more realism errors. In fact, however, deceptiveness had little effect on AR performance.

Another common account attributes children's AR errors to an immature theory of mind. Presumably, young children fail to represent their own changed belief about a deceptive object's identity (i.e., function) after it is revealed, and this somehow elicits perseverative errors. By this account, AR, false belief, and perspective-taking errors stem from a single conceptual limitation. This explanation is not, however, empirically well supported. Gopnik and Astington (1988) found that object AR and false-belief tasks shared about 20% of variance across children, but they did not control any confounds such as age, verbal knowledge, memory, or test materials. Frye et al. (1995) did control for age and found that AR and false belief tests shared 18% of variance; other studies, however, found no more than 10% shared variance between AR and false belief or perspective taking tasks (Miller, Holmes, Gitten, & Danbury, 1997; Ray, 1996). On average, then, a small portion of AR variance is shared by theory of mind tasks even when age, and no other confounding factor, is controlled. Though we did not test children's ability to infer false beliefs, it is notable that our control tests—which do not require inferences about mental states, beliefs, or perspectives—did correlate significantly with AR performance even when age, vocabulary, and memory span were controlled.

These findings lead us to believe that AR errors do not depend on deceptive objects, predicates about perception (e.g., "looks to your eyes like a …"), or inference about mental states. Instead, we offer the following empirically defensible, relatively parsimonious description of the proposed tendency of 3- to 5-year-old children to perseverate in paradigms like the AR test:

- If making two or more forced-choice responses between verbal options (e.g., category labels), with respect to a complex stimulus,
- and each response follows a discrete question about the stimulus,
- and it is not obvious that each question calls for an independent response (e.g., questions are somewhat ambiguous vis-à-vis response choices or labels),
- then select one response for that stimulus and repeat it (i.e., perseverate) until corrective feedback is given or a new question clearly implies a different response option.

This tendency also can describe children's performance in paradigms including deductive card sorting, inductive word learning, and other tasks (Deák, 2000; Welsh, Pennington, & Groisser, 1991; Zelazo, Frye, & Rapus, 1996) including the AR and control tasks. This description also sharpens Siegal's (1991) claims that the pragmatics of the AR test are odd to children. We agree that young children, when answering hard questions, are motivated to keep the adult happy, not to make semantically exacting interpretations. The previous description, however, also specifies (to some extent at least) the discourse conditions that elicit perseverative responses. Note also that the description is not incompatible with Zelazo and Frye's (1998) cognitive complexity and control (CCC) account of children's perseveration. That is, the process outlined previously might be elicited more readily when response contingencies are complex, as CCC theory stipulates.

The previous description does not fully explain why children respond by perseverating, and because some of the evidence at hand is correlational, we cannot draw a complete causal account. We believe, however, that some clues can be gleaned from the data described here.

Receptive vocabulary significantly predicts children's fluency in naming multiple aspects of complex entities (e.g., objects, characters; Deák & Maratsos, 1998; Deák et al., 2001). Though standardized receptive vocabulary tests (e.g., PPVT–R) directly assess shallow lexical knowledge, they predict breadth and accessibility of semantic knowledge of words and predicates. Vocabulary is a product of a child's history of semantic inferences about his or her language, and it differs widely across children. It determines the efficiency and specificity with which a child can select appropriate descriptions (e.g., noun phrases) for an aspect of an entity or situation. It therefore can be used to predict whether a child is likely to map specific questions to specific, appropriate phrases or labels. Thus, predicate mapping is directly related to performance in the AR and other tests, and it is partly determined by receptive vocabulary. (Of course, vocabulary predicts full-scale IQ, and it is possible that vocabulary and AR performance in preschool children both vary with overall IQ.)

Nonspecific predicate mapping does not invariably cause perseveration: A child might instead guess at the answer, ask for additional information, and so on. In fact, a few children (14% and 5% in Experiments 1 and 2, respectively) seemed to alternate answers to successive questions, apparently at random, thereby swinging between correct and switched responses. It makes sense that some children, when challenged by the discourse and verbal factors outlined earlier, would respond by switching responses instead of perseverating. Still, perseveration is the most prevalent error pattern, and we wish to know why.

Perseveration might be related to the logical skill of detecting indeterminacy across questions. Children who are insensitive to the indeterminacy of each successive question would see no reason to alter their first responses, if made with confidence. Yet the relation of AR errors to indeterminacy detection remains ambiguous, in part because correlational data reveal nothing about the source of the association. Notably, errors in the indeterminacy detection test also are perseverative, and combining these errors with control tests perseverative errors (Experiment 2) yielded a good predictor of AR performance. Thus, the simpler interpretation is that the perseverative tendency, rather than detection of indeterminacy, explains the correlation between the AR and indeterminacy detection test. Also, the hidden exemplar condition (Experiment 2), which eliminated concrete cues to verbal responses but might have highlighted indeterminacy by making the choices concrete, actually increased AR perseverative errors, contrary to the indeterminacy insensitivity hypothesis. Thus, the data do not show a relation between children's sensitivity to indeterminacy and their tendency to perseverate across questions.

Other questions about perseveration merit further research. For example, it is often assumed that perseveration stems from failure to inhibit prior responses, due to immature inhibitory mechanisms in frontal cortex. Yet, recent work suggests that perseveration is not always due to inhibitory failure (Deák & Narasimham, 2003; Jacques, Zelazo, Kirkham, & Semecesen, 1999); therefore, perhaps children's AR errors do not reflect failure to inhibit the first label produced. Consistent with this possibility, many children in Experiment 1, when shown a control test item, labeled both aspects (e.g., ''It's a duck wearing a hat!''), then went on to perseverate when answering questions. Apparently perseverative children are not bound to repeat the first label produced, or they could not have initially produced both labels, spontaneously and in rapid succession.

If children possess the conceptual flexibility to switch labels but make many errors in the AR test and other similar tests, what are the implications? A pessimistic interpretation is that our field has simply been mislead by an artifact of the discourse structure of certain experimental paradigms, including the AR test. By this reading, AR errors are a laboratory phenomenon that have little bearing on preschool children's everyday conceptual abilities. However, children's perseverative errors are definable and predictable, and experimental questions are not completely dissimilar from other questions posed to children in everyday settings such as schools. In fact, the discourse structure that evokes perseveration can be loosely characterized as ''classroom questioning,'' a sort of interaction that happens frequently in schools, even in preschools. The implication is that verbal questioning of preschoolers by teachers might elicit systematic errors. Consequently, teachers of young children might learn what kinds of questioning patterns elicit errors and devise other modes of questioning preschool children. For example, the label elicitation task of Deák and Maratsos (1998) is similar in purpose and content to the AR test, yet it reveals competence in children's conceptual and labeling flexibility. It would be useful to observe how often teachers and other professionals (e.g., pediatricians) ask preschool children series of forced-choice questions and whether this leads the adult to under-

estimate children's conceptual grasp of the question content.

In sum, two verbal factors account for up to half of the variability in preschool children's ability to label correctly the intended appearance and function of deceptive objects. A general tendency to perseverate over responses to forced-choice questions explains, at least in part, perseverative errors in a range of tasks with the same discourse structure (but different materials and questions) as the AR test. Also, receptive vocabulary, which reflects verbal inference and predicate mapping skills, mediates children's flexible selection of multiple labels for a complex entity. In contrast, verbal working memory does not uniquely predict a significant portion of children's AR performance variance, and manipulation of perceptual cues to object identity (e.g., object deceptiveness, showing exemplars before the test questions) produces paradoxical or nonsignificant effects. There was no convincing evidence that ability to detect the indeterminacy of a question plays a unique role in children's changing AR abilities, though this remains a possibility for future research. Most notable is the finding that factors long believed to define children's AR performance—object deceptiveness and the conceptual ability to hold dual representations of object identity—are peripheral to AR errors, whereas discourse understanding and verbal knowledge seem to be central.

References

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (pp. 96–151). Oxford, England: Blackwell.

Brenneman, K., & Gelman, R. (1993, March). *Reasoning about object identities in the appearance-reality situation.* Presented at the biennial SRCD meeting, New Orleans, LA.

Campbell, R. N., & Bowe, T. B. (1983). Text and context in early language comprehension. In M. Donaldson, R. Grieve, & C. Pratt (Eds.), *Early childhood development and education* (pp. 115–126). Oxford, England: Blackwell.

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

Cosgrove, J. M., & Patterson, C. J. (1977). Plans and the development of listener skills. *Developmental Psychology, 13,* 557–564.

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

Deák, G. O. & 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 flexible inferences about word meanings*. Manuscript submitted for publication.

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.

DeLoache, J. (1995). Early symbol understanding and use. In D. L. Medin (Ed.), *The psychology of learning and motivation* (Vol. 33, pp. 65–113). San Diego, CA: Academic Press.

Dempster, F. N. (1992). The rise and fall of the inhibitory mechanism: Toward a unified theory of cognitive development and aging. *Developmental Review, 12,* 45–75.

De Vries, R. (1969). Constancy of generic identity in the years three to six. *Monographs of the Society for Research on Child Development, 34*(3, Serial No. 127).

Donaldson, M. (1978). *Children's minds.* Glasgow, Scotland: Fontana.

Dunn, L. M., & Dunn, L. (1981). *Peabody Picture Vocabulary Test–Revised*. Circle Pines, MN: American Guidance Service.

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

Flavell, J. H., Flavell, E. R., & Green, F. L. (1987). Young children's knowledge about the apparent-real and pretend-real distinctions. *Developmental Psychology, 23,* 816–822.

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., Green, F., & Flavell, E. (1989). Young children's ability to differentiate appearance-reality and level 2 perspectives in the tactile modality. *Child Development, 60,* 201–213.

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. (2000). Theory of mind, domain specificity, and reasoning. In P. Mitchell & K. Riggs (Eds.), *Children's reasoning and the mind* (pp. 149–167). Hove, England: Psychology Press.

Frye, D., Zelazo, P. D., & Palfai, T. (1995). 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., 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.
Gauvain, M., & Greene, J. K. (1994). What do young children know about objects? *Cognitive Development, 9,* 311–329.
Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review, 62,* 32–41.
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.
Harris, P. L., Donnelly, K., Guz, G. R., & Pitt-Watson, R. (1986). Children's understanding of the distinction between real and apparent emotion. *Child Development, 57,* 895–909.
Harris, P. L., & Leevers, H. J. (2000). Reasoning from false premises. In P. Mitchell & K. Riggs (Eds.), *Children's reasoning and the mind* (pp. 67–86). Hove, England: Psychology Press.
Houdé, O. (2000). Inhibition and cognitive development: Object, number, categorization, and reasoning. *Cognitive Development, 15,* 63–73.
Huttenlocher, J., Haight, W., Bryk, A., Seltzer, M., & Lyons, T. (1991). Early vocabulary growth: Relation to language input and gender. *Developmental Psychology, 27,* 236–248.
Jacques, S., Zelazo, P. D., Kirkham, N., & Semecesen, T. (1999). Rule selection vs. rule execution in preschoolers: An error-detection approach. *Developmental Psychology, 35,* 770–780.
Lakoff, G. (1973). Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of Philosophical Logic, 2,* 458–508.
Markman, E. M. (1979). Realizing that you don't understand: Elementary school children's awareness of inconsistencies. *Child Development, 50,* 643–655.
Merriman, W. E., & Bowman, L. (1989). The mutual exclusivity bias in children's word learning. *Monographs of the Society for Research on Child Development. 54*(3-4, Serial No. 220).
Merriman, W. E., Jarvis, L. H., & Marazita, J. M. (1995). How shall a deceptive thing be called? *Journal of Child Language, 22,* 129–149.
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.
Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development, 60,* 167–171.
Nagy, W. E., & Herman, P. A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. McKeown & M. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19–35). Hillsdale, NJ: Erlbaum.
Olson, D. R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review, 47,* 257–281.
Patterson, C. J., Cosgrove, J. M., & O'Brien, R. G. (1980). Nonverbal indicants of comprehension and noncomprehension in children. *Developmental Psychology, 16,* 38–48.
Ray, S. D. (1996). *The relation between appearance-reality and perspective taking in young children.* Unpublished master's thesis, Middle Tennessee State University, Murfreesboro, TN.
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.
Sattler, J. M. (1988). *Assessment of children* (3rd ed.). San Diego, CA: Sattler.
Siegal, M. (1991). *Knowing children: Experiments in conversation and cognition.* Hillsdale, NJ: Erlbaum.
Speer, J. R. (1984). Two practical strategies young children use to interpret vague instructions. *Child Development, 55,* 1811–1819.
Taylor, M., & Hort, B. (1990). Can children be trained in making the distinction between appearance and reality? *Cognitive Development, 5,* 89–99.
Wellman, H. M., & Estes, D. (1986). Early understanding of mental entities: A reexamination of childhood realism. *Child Development, 57,* 910–923.
Welsh, M. C., Pennington, B. F., & Groisser, D. B. (1991). A normative-developmental study of executive function: A window on prefrontal function in children. *Developmental Neuropsychology, 7,* 131–149.
Woolley, J. D., & Wellman, H. M. (1990). Young children's understanding of realities, nonrealities, and appearances. *Child Development, 61,* 946–961.
Zelazo, P. D., & Frye, D. (1998). Cognitive complexity and control: II. The development of executive function in childhood. *Current Directions in Psychological Science, 7,* 121–126.
Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development, 11,* 37–63.

Appendix

AR Task Protocol (Experiment 1)

Demonstration (sample discourse): What does this [rock box] look like to you? [That's right/Actually] it looks like a rock. But really and truly it's not a rock … it is a hiding place. See? You can open it and put pennies inside. What is it really and truly? [That's right/Actually] its a hiding place, but it looks like a rock. [Note: Exemplars condition compared object with exemplars.]

Test: What is this really and truly? Is it really and truly a rock or is it really and truly a box? When you look at this right now, does it look like a box or does it look like a rock?

Control Task Protocol (Experiment 1)

Demonstration: What does this look like to you right now? [That's right/Actually] this looks like a bird. But it is not just a bird. This bird is holding something. It has a crayon. What does it have? [That's right/Actually] it has a crayon. It looks like a bird and it has a crayon.

Test: What does it have? Does it have a bird or does it have a crayon? What does it look like? Does it look like a crayon or does it look like a bird?

AR Test (Experiment 2): Standard Questions

Demonstration (sample discourse): What does this [car crayon] look like? [That's right/Actually] it looks like a car. But really and truly it is not a car. Really and truly it is a crayon. See? You can color with it [demonstrate function]. Now, what is this really and truly? [That's right/Actually] it's…crayon. But…it looks like a car.

Test: What is this really and truly? Is it really and truly a crayon or is it really and truly a car? When you look at this, does it look like a car or does it look like a crayon?

AR Test (Experiment 2): Modified Questions

Demonstration: [For example:] Look at this… its color and shape. What does it look like? … Let's see what you really … do with it. [demonstrate function] See, you really use it like ….

Test: What are you really supposed to do with this? Do you really use it as a car or … a crayon? When you look at it, is its color and shape like a car or…a crayon?

Control Test: Overlap (Includes "Looks Like")

Demonstration: What does this look like? [That's right/Actually] it looks like a bear. And what does this have? [That's right/Actually], it has a key. So, what does this look like? Does it look like a bear or does it look like a key? What does this have? Does it have a bear or does it have a key?

Test: What does this look like? Does it look like a key or does it look like a bear? What does it have? Does it have a key or does it have a bear?

Control Test: Nonoverlap (Includes "Is Made of")

Demonstration: What is this made of? [That's right/Actually] it's made of Play Doh. And what does this have? [That's right/Actually] it has a key. So, what is it made of? Is it made of Play Doh or is it made of a key? What does it have? Does it have Play Doh or does it have a key?

Test: What is this made of? Is it made of Play Doh or is it made of a key? What does it have? Does it have Play Doh or does it have a key?

Detection of Indeterminacy Test (Example)

Demonstration: [Show both spinners] See how these work? You can spin these. See this one [determinate]? What are the pictures? [Yes/Actually] they're all Winnie the Pooh! Are they all the same or … different? [That's right/Actually] they're all the same! See this one [indeterminate] …. Are they all the same or…different? [Yes/Actually] they are all different!

Test: I am going to spin it [first spinner]. Right now, before I spin it, do you have to guess who it will point to when it stops, or do you know for sure?