Cognitive Psychology 99 (2017) 17–43

Contents lists available at ScienceDirect

Cognitive Psychology

journal homepage: www.elsevier.com/locate/cogpsych

Children's representation of abstract relations in relational/array match-to-sample tasks

Jean-Rémy Hochmann[a,b,*], Arin S. Tuerk[c], Sophia Sanborn[d], Rebecca Zhu[c], Robert Long[e], Meg Dempster[f], Susan Carey[c]

[a] CNRS, UMR 5304, Institut des Sciences Cognitives - Marc Jeannerod, 67 Bd Pinel, 69675 Bron, France
[b] Université Claude Bernard Lyon 1, France
[c] Department of Psychology, Harvard University, William James Hall, 33 Kirkland Street, Cambridge, MA 02138, USA
[d] Department of Psychology, UC Berkeley, Tolman Hall, Berkeley, CA 94720-1650, USA
[e] Department of Philosophy, New York University, 5 Washington Place, New York, NY 10003, USA
[f] Department of Psychology, University of Bath, Claverton Down, Bath, United Kingdom

ABSTRACT

Five experiments compared preschool children's performance to that of adults and of non-human animals on match to sample tasks involving 2-item or 16-item arrays that varied according to their composition of same or different items (Array Match-to-Sample, AMTS). They establish that, like non-human animals in most studies, 3- and 4-year-olds fail 2-item AMTS (the classic relational match to sample task introduced into the literature by Premack, 1983), and that robust success is not observed until age 6. They also establish that 3-year-olds, like non-human animal species, succeed only when they are able to encode stimuli in terms of entropy, a property of an array (namely its internal variability), rather than relations among the individuals in the array (same vs. different), whereas adults solve both 2-item and 16-item AMTS on the basis of the relations same and different. As in the case of non-human animals, the acuity of 3- and 4-year-olds' representation of entropy is insufficient to solve the 2-item same-different AMTS task. At age 4, behavior begins to contrast with that of non-human species. On 16-item AMTS, a subgroup of 4-year-olds induce a categorical rule matching all-same arrays to all-same arrays, while matching other arrays (mixed arrays of same and different items) to all-different arrays. These children tend to justify their choices using the words "same" and "different." By age 4 a number of our participants succeed at 2-item AMTS, also justifying their choices by explicit verbal appeals using words for same and different. Taken together these results suggest that the recruitment of the relational representations corresponding to the meaning of these words contributes to the better performance over the preschool years at solving array match-to-sample tasks.

1. Introduction

It is uncontroversial that the cognition of human adults differs from that of other animals in its ability to represent abstract and combinatorial concepts (see, for example, Penn, Holyoak, & Povinelli, 2008). Human adults represent abstract concepts such as *freedom* and *atom*, and they can combine concepts to produce novel ones such as *oxygen atom* and *not an atom*. If these abilities are present in other animals, they are certainly not present to the same extent as in human adults. What is still unknown is whether the

* Corresponding author at: CNRS, UMR5304, Institut des Sciences Cognitives - Marc Jeannerod, 67 Bd Pinel, 69675 Bron, France.
E-mail address: hochmann@isc.cnrs.fr (J.-R. Hochmann).

https://doi.org/10.1016/j.cogpsych.2017.11.001
Accepted 3 November 2017
Available online 10 November 2017
0010-0285/

ability to represent abstract combinatorial concepts is inherent to human nature, attested in prelinguistic infants' thought, or whether some learning processes, perhaps involving language acquisition, give rise to these abilities later in childhood. Some have argued that the development of syntax and/or the acquisition of the lexicon could play a role in the ontogenesis of uniquely human abstract, combinatorial, cognitive abilities (Gentner, 2003; Spelke, 2003; Waxman & Braun, 2005).

Exploring the evolutionary and ontogenetic origins of the concepts *same* and *different* is a good case study to address these issues (see Penn et al., 2008). Representations of these relations are abstract because they generalize across modalities and domains. Two images can be the same, as well as two sounds, two smells, or two ideas (e.g., *freedom* and *liberty*). The format of representations of same and different cannot be simply iconic, and these representations themselves exemplify combinatoriality: *different* is *not same* and *same* is *not different*. Furthermore, *same* and *different* are relational concepts, perhaps the most fundamental relations in thought (James, 1890/1950), and thus participate in *relational reasoning*, which in turn plays an important role in human conceptual development (Carey, 2009; Gentner, 2003; Penn et al., 2008).

Many studies, from three different paradigms (match to sample, same/different discrimination rule learning, relational match to sample) suggest that a wide variety of animal species, from bees to chimps, can condition *behaviors* on sameness and difference (e.g., Giurfa, Zhang, Jenett, Menzel, & Srinivasan, 2001; Harley, Putman, & Roitblat, 2003; Mumby, 2001; Thompson & Oden, 1995; Wasserman & Young, 2010; Wright, Cook, Rivera, Sands, & Delius, 1988), and young infants, years before they learn the words "same" and different", also have these capacities (Hochmann, Benavides-Varela, Fló, Nespor, & Mehler, 2017; Hochmann, Benavides-Varela, Nespor, & Mehler, 2011; Hochmann, Carey, & Mehler, submitted for publication; Hochmann, Mody, & Carey, 2016; Kovács, 2014). But for each behavior that reflects representations of same and/or different, a number of questions arise. First, is there some other way the animal or infant could be passing the task, other than on the basis of representations with the content *same* or *different*? Second, does the task require internal representations with the content *same* and *different* that go beyond matching computations that are largely spread throughout the animal kingdom and involved in all recognition and categorization processes. Third, even if there is good evidence that success is based on the concepts *same* and/or *different*, are there evolutionary or ontogenetic changes in the nature of the representations of these relations and in the computations they support?

1.1. Match to sample tasks

The earliest data that were taken as evidence for non-linguistic representations of both same and different derived from the capacity of animals and infants to solve the match-to-sample and the non-match-to-sample tasks (bees: Giurfa et al., 2001; pigeons: Blough, 1959; Wright et al., 1988; dolphins: Harley et al., 2003; rats: Mumby, 2001; apes: Oden, Thompson, & Premack, 1988; human infants: Hochmann et al., 2016). In the match-to-sample (MTS) task, participants must learn to choose, between two possible choices, the stimulus that is the same as the sample (e.g., between choice stimuli consisting of a square and a triangle, if the sample is a square choose the other square) and generalize the rule to novel stimuli. In the non-match-to-sample task (NMTS), the rule is to choose the stimulus that is different from the sample (i.e., in the above example, if the sample is the square, choose the triangle). Because the rule learned generalizes freely to novel stimuli, it is likely that some representations of same and/or different underlie success. Zentall, Edwards, Moore, and Hogan (1981) and Hochmann et al. (2016) provide evidence that performance on *both* MTS and NMTS relies on the representation *same* alone, and suggest that the representation *same* may be *entirely implicit*, carried by a match computation. That is, the procedure infants and animals may be using in MTS could be: place representation of sample in working memory: *x*; subsequently, if encounter *x*, select *x*, whereas in NMTS, the procedure might be: place representation of sample in working memory: *x*; subsequently, if encounter *x*, avoid *x*. The abstractness in these procedures is carried by lack of constraints on the content of the variable *x*. What "implicit" means in this context is that there is no mental symbol for the relation same in this procedure: the content *same* is carried by the match computation that is involved in all recognition and categorization processes.

1.2. Same/different discrimination tasks

Even if the above proposal is correct for the representations that underlie MTS and NMTS, both infants and non-human animals *also* apparently have the capacity to create mental symbols for the relation same, and perhaps also for the relation different. Both animals and infants can solve same/different discrimination tasks, which at face-value involve learning rules formulated in terms of the relation between individuals. That is, they can apparently learn rules such as "if same, choose red; if different, choose green" (for animals, see Thompson & Oden, 1995; Wasserman & Young, 2010 for reviews). Infants can learn rules with the content *if the two items that appear in the middle are the same, look right* (Hochmann, 2010; Hochmann et al., 2011; Hochmann et al., 2017; Hochmann et al., submitted for publication; Kovács, 2014) or *to activate the blicket detector, choose the pair of objects that are the same* (Walker, Bridgers, & Gopnik, 2016; Walker & Gopnik, 2014).[1]

Animals and infants could possibly succeed on same/different discrimination tasks not on the basis of the relation between the individuals in a pair of stimuli, but rather on some global property of the pair, such as symmetry or the degree of variability among the elements in an array (a property that can be modelled by the entropy measure, see below). Indeed, a number of animal species

[1] Infants also can be habituated to pairs of stimuli that instantiate the relations same or different, recovering interest when presented with a pair that instantiates the opposite relation (Addyman & Mareschal, 2010; Ferry, Hespos, & Gentner, 2015; Tyrrell, Stauffer, & Snowman, 1991). We do not discuss the habituation results here, for we are concerned with representations that are the input to behavioral choices. However, habituation studies certainly implicate some representations of same and/or different, and the questions we raise here arise for these studies as well. Is there some other basis of response (e.g., symmetry) in these studies, and if not, what is the nature of the representations of same and/or different that underlies habituation?

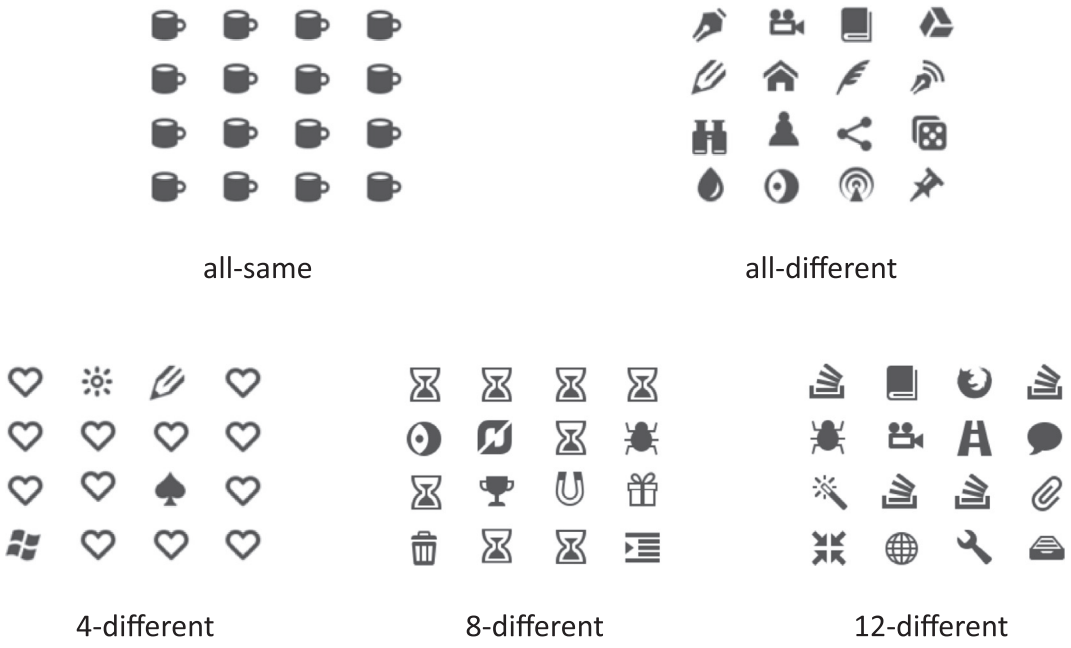

Fig. 1. Examples of 16-item all-same and all-different and intermediary arrays.

have been shown to be able to learn behaviors conditioned by the presence or absence of symmetry in a visual display (pigeons: Delius & Nowak, 1982; bees: Giurfa, Eichmann, & Menzel, 1996; dolphins: von Fersen, Manos, Goldowsky, & Roitblat, 1992; sharks: Schluessel, Beil, Weber, & Bleckmann, 2014). Moreover, in paradigms where animals tend to fail at learning rules requiring discriminating *pairs* of same or different elements, both pigeons and baboons *can* learn rules conditioned on the distinction between arrays with 16 identical elements and 16 elements all different from each other (Wasserman, Young, & Fagot, 2001; Young, Wasserman, & Garner, 1997; see Fig. 1 for examples of such arrays). This result is consistent with the possibility that 16-item all-same or all-different arrays make the relations same and different more salient than do 2-item both-same or both-different arrays. However, several results establish that animals' behavior in these large array tasks is instead controlled by the computation of entropy measuring the variability of items within an array (see Wasserman and Young (2010), for review).

The entropy of an array is a function of the number of same and different elements composing the array, so that the larger the number of different elements, the higher the entropy, and an array of same elements has, by definition, null entropy. Borrowed from information theory, entropy models the number of bits necessary to encode the different elements of an array. The spatial arrangement of the elements of the array does not, by definition, affect the computation of entropy. Following Wasserman and Young (2010), in the present work, we compute the entropy H of an array A with the following equation: $H(A) = -d/N \log(1/N) - s/N \log(s/N)$, where N is the total number of elements composing the array A, d the number of these elements that are different and s the number of elements that are the same. Of course, as should be apparent in the equation, the computation of entropy relies on computations of sameness and differences. It is important to stress that these computations do not constitute representations of the abstract relations same and different, but rather local low-level perceptual computations.

Entropy is a dimension on which arrays vary continuously, like approximate cardinal value. Such dimensions of variation, whether of arrays or individuals, are encoded by analog symbols that are a logarithmic function of their real-world value, such that discrimination is governed by Weber's law (Fechner, 1966/1860). This law, initially enunciated for perception, dictates that the change in a stimulus that will be just noticeable is a constant ratio of the original stimulus. In consequence, the same magnitude of variation should be more noticeable in lower than in higher ranges of the dimension. Fig. 2 displays idealized versions of the four most common patterns of classification observed in same-different discrimination studies. If entropy is guiding animals' choice, the probability that intermediate sample arrays (e.g., 4 same-12 different arrays or 8 same-8 different arrays; see Fig. 1) are classified with the all-different choice arrays should increase logarithmically with increasing entropy of the sample array. This is what is observed in the animal array discrimination studies, after learning the discrimination of 16-same from 16-different arrays (Wasserman, Fagot, & Young, 2001; Wasserman, Young, & Nolan, 2000; Young & Wasserman, 1997; see Fig. 2a). This pattern contrasts with the behavior of the majority (about 80%) of human adults given the same training as animals. Most adults induce a categorical distinction between all-same vs. not-all-same from this training, and categorize all arrays in which the items are not all the same with the all-different response (Castro & Wasserman, 2013; Castro, Young, & Wasserman, 2006; Young & Wasserman, 2001; see Fig. 2b). Some adults induce a categorical distinction between all-different and not-all different, and classify intermediate arrays as in Fig. 2c (e.g., Castro & Wasserman, 2013). Wasserman and his colleagues systematically separate these two categorical patterns of classification (Fig. 2b and c) from what they call "continuous patterns," and take continuous patterns as reflecting entropy discrimination. Both pattern 2a and 2d are thus called "continuous patterns." But the actual continuous patterns in human adult

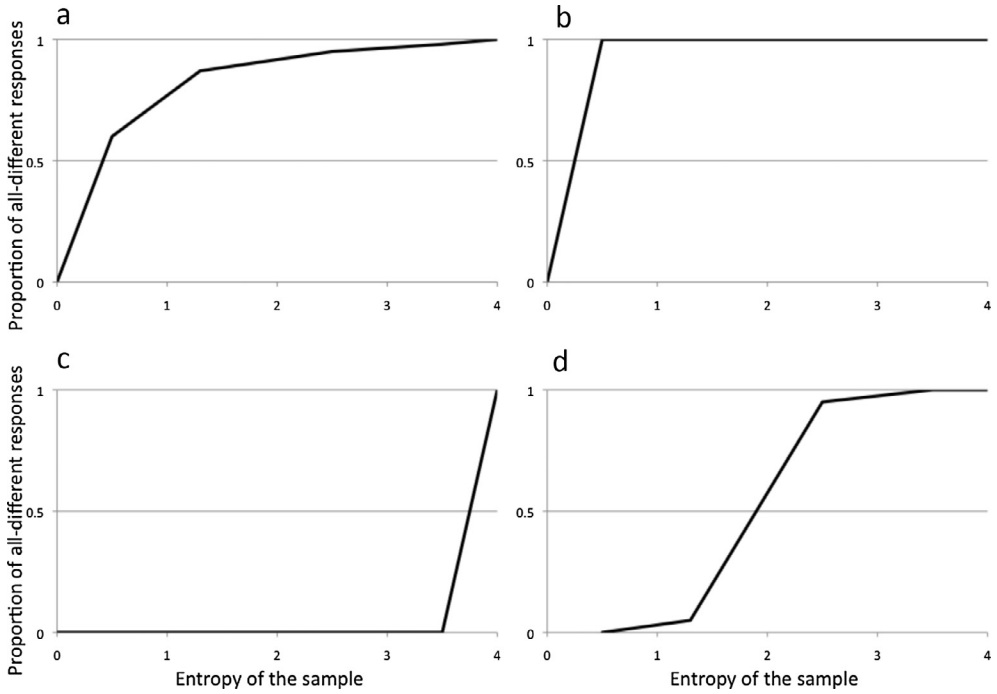

Fig. 2. Four idealized patterns of responses in same-different discrimination tasks.

discrimination learning tasks almost always approximate that of Fig. 2d, in which choices are decidedly not a logarithmic function of the entropy of the intermediate arrays. This pattern most likely reflects having learned yet another categorical rule, classifying arrays as all-same and all-different arrays, a rule that cannot be applied to the intermediate arrays. Choices in test then are based on similarity of each sample to either the all-same array and the all-different-array. Here we distinguish all four patterns of classification, labelled as on Fig. 2.

Wasserman and his colleagues are well aware, of course, that the finding of graded responses on intermediate arrays, such that the probability of the all-different response increases monotonically as a function of increasing entropy of the stimulus being classified, does not conclusively show that animals are basing their response on entropy. Therefore, two further analytic approaches were adopted to adjudicate between responses based on entropy or the relations same and different: first, an exploration of training transfer to arrays of different sizes, and second, a comparison of performance levels on trials where arrays have different compositions but share a given entropy level.

A representation such as *all same* applies to 16-item, 8-item, 4-item and even 2-item arrays of identical elements, and a representation such as *all different* similarly applies to 16-item, 8-item, 4-item and even 2-item arrays of unique elements. But the entropy differences between the all-same and all-different arrays vary widely as a function of array sizes (e.g., for 16-item arrays the contrast is between 0 and 4; for the 2-item arrays the contrast is between 0 and 1). For pigeons and baboons, successful performance on 16-item arrays (16 same vs. 16 different) transfers to 8-item arrays (discriminating 8 same vs. 8 different), but not to 2-item arrays (discriminating 2 same vs. 2 different; Wasserman et al., 2001; Young et al., 1997). Since the relations same and different are defined over all those array sizes, this failure of transfer suggests that the rules learned were not articulated in terms of these relations.

Further evidence that entropy is driving categorization is that arrays with a given entropy level are treated alike no matter how that entropy level is achieved (e.g., for 16-item arrays, a mixture of 8 pairs of items – 2a-2b-2c-2d-2e-2f-2g-2h – has entropy 3, as do the following mixtures: 5a-3b-1c-1d-1e-1f-1g-1h-1i-1j, 6a-1b-1c-1d-1e-1f-1g-1h-1i-1j-1k; Castro & Wasserman, 2013)- as is the case in the animal studies and in some of the human "continuous" patterns (Wasserman and colleagues' designation of any non-categorical patterns, such as those depicted on Fig. 2a and d; Castro & Wasserman, 2013). For example, pigeons tend to give the "same" response to 2-item different pairs (entropy 1), just as much as they do for 2-item same pairs (entropy 0; Young et al., 1997). Similarly, within 16-item arrays, both pigeons and baboons appear to have difficulty discriminating between 13 same-3 different arrays (entropy 1) and 16 same arrays (entropy 0) (Wasserman et al., 2001; Young & Wasserman, 1997). Taken together, these patterns of performance on tasks of training transfer and discriminability of various arrays of a given entropy provide strong evidence that pigeons' and baboons' success in discriminating 16-item all-same (entropy 0) arrays from 16-item all-different (entropy 4) arrays is based on entropy, not on the relations same and different.

Thus, there is no doubt that on many tasks in which animals are conditioned to respond to arrays of elements which are the same as each other, contrasting with arrays where the elements are different from each other, performance is often guided by entropy representations. These representations reflect a global property of the array (degree of variability), and not the relations among elements within the array. In contrast, the majority of human adults, who are trained on 16-item discrimination, base their choices on

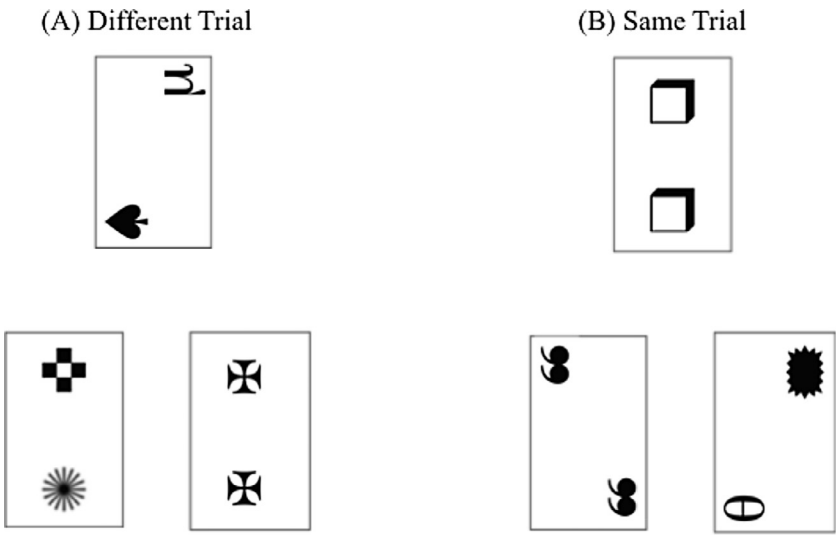

Fig. 3. Examples of different-trial (A) and same-trial (B) in a relational match-to-sample.

the categorical contrast between all-same and not all-same.

Not noted above was the fact that a non-negligible percentage of adult human participants fail to learn the discrimination *at all* under this training regime (8% percent in Young & Wasserman, 2001, 21% in Castro & Wasserman, 2013). Castro and Wasserman (2013) also included a group of participants who were initially trained to discriminate 2-item arrays (2 same vs. 2 different). An astonishing 52% of adults failed to reach criterion within 48 training trials, and *none* of those who did so induced a rule based on graded entropy. Rather, they either sorted test arrays according to the contrast all-same vs. not all-same (Fig. 2b, 69%), or the contrast all-different vs. not all-different (Fig. 2c, 10%), or the contrast between all-same and all-different (Fig. 2d), treating the intermediate arrays according to whether they were more similar to the all-same-choice or the all-different choice (21%). These results support two conclusions. First, even for adults the rule discrimination task does not always elicit representations based on *either* representations of the relations same and different or of the property of entropy. That is, participants who failed to learn the rule at all did not consistently appeal to entropy either. Second, 2-item arrays elicit quite different spontaneous representations than do 16-item arrays.

1.3. Relational match to sample

In an attempt to seek evidence for representations of the relations same and different, as well as to explore whether non-human primates were capable of analogical reasoning, Premack (1983) designed the *relational match to sample task (RMTS),* in which pairs of stimuli are presented as choices, one exemplifying the relation same (e.g., A A) and the other the relation different (e.g., B C). In relational match-to sample, if the sample is X X, the correct choice is A A rather than B C; if the sample is X Y, the correct choice is B C rather than A A (see Fig. 3). Premack reasoned that, contrary to the simple MTS and NMTS (see above), the relational match-to-sample requires summary mental symbols that mean both same and different, for the sample card would plausibly be represented by a symbol that specified the relation (e.g. same) and each of the choice cards by symbols *same* or *different* to support the decision of which choice card constitutes a match. Furthermore, this task can be solved on the basis of a structure mapping analogy: the correct choice instantiates the same relation between the two items in the array, rather than an object feature or object identity match, making it an assay of analogical reasoning. In his initial studies, Premack found that his language trained chimp, Sarah, who had been taught summary symbols that meant “same” and “different” succeeded at RMTS, whereas non-language trained chimps did not. Subsequent work found that apes previously trained on a conditional same/different discrimination task could succeed at RMTS, but those without this previous training could not (see also Thompson, Oden, & Boysen, 1997). Further work has confirmed the extreme difficulty of RMTS for animals (Thompson & Oden, 1995; Wasserman & Young, 2010) and has been taken as evidence that animals might not have summary symbols for the relations same and different, and/or might not be capable of analogical reasoning (Penn et al., 2008).

Furthermore, research from Wasserman’s group has confirmed that when animals solve RMTS tasks, they often do so on the basis of entropy (matching high entropy arrays with high entropy arrays and low entropy arrays with low entropy arrays) rather than the relations same and different (Cook & Wasserman, 2007; Fagot, Wasserman, & Young, 2001; Wasserman & Young, 2010). For this reason, in this paper we will refer to this task as Array Match to Sample (AMTS), rather than RMTS, so as not to prejudge the basis for success on it. The classic RMTS, as designed by Premack (1983), corresponds to a 2-item AMTS, while the modified version with 16 pictures (Cook & Wasserman, 2007; Fagot et al., 2001) corresponds to a 16-item AMTS.

While more recent work shows that animals *can* solve 2-item AMTS (Fagot & Parron, 2010; Fagot & Thompson, 2011), these successes could also be based on global properties of the array, like symmetry or entropy. In these studies, *massive* training (e.g., over

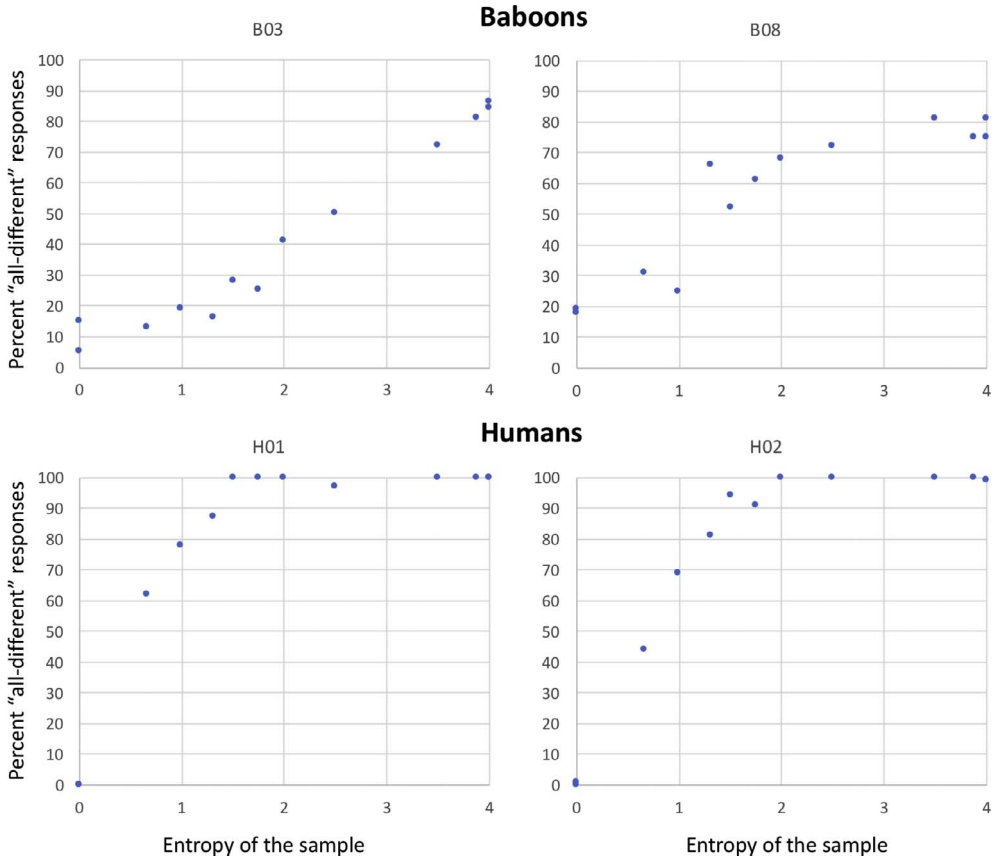

Fig. 4. Performance of two baboons and two humans on a 16-item AMTS (Experiment 3 from Fagot, Wasserman & Young, 2001).

35,000 training trials) was required, which may have had the effect of tuning entropy acuity so that the contrast between entropy 0 and entropy 1 was discriminable, or allowing other types of stimulus property to drive choice. Some recent studies, however, found more rapid success on 2-item AMTS (e.g., Obozova, Smirnova, Zorina, & Wasserman, 2015; Smirnova, Zorina, Obozova, & Wasserman, 2015; Vonk, 2003). We will return to these latter studies in the general discussion.

Only one study we know of (Fagot et al., 2001) has compared animals (2 baboons) with human adults (2 humans) on 16-item AMTS, in a study in which, after learning to match 16-same samples with 16-same choices and 16-different samples with 16-different choices, they were subsequently tested on intermediate sample arrays to explore the basis of their success. Both baboons succeeded on the task, and did so on the basis of a graded response to entropy. Of course, both adults also succeeded, but perhaps surprisingly, also apparently did so on the basis of entropy. Fig. 4 (adapted from Fagot et al., 2001) shows that for both the baboons and the humans, the percentage of selection of the 16-different stimulus is a logarithmic function of the entropy of the target card. One goal of Experiment 1 is to systematically explore human adults' performance on 16-item AMTS, under the same testing protocol we will subsequently use with preschool children, expanding the published data from adults tested on this task from 2 to 100.

1.4. Appealing to developmental data

Two-item AMTS appears to be hard for young children, just as it is for non-human animals. In a review paper on his seminal work, Premack reports: "Children below about 4 years of age failed both the XX-AA and the XY-CD type of problems; by about 4½ they passed XX-AA but still failed XY-CD; and they may have to be as old as 6 before passing both types of problems (Premack & McClure, in preparation; [cited in Premack, 1983; p. 128])". Unfortunately, thirty-five years later the Premack & McClure article still awaits publication. Surprisingly, since then, no report has been published looking at children's performance in this task. Christie and Gentner (2014) tested children on a XX-AA problem that could be solved by learning a single rule "choose same;" i.e., this could be a discrimination learning task, not an AMTS task. Consistent with Premack and McClure, they found spontaneous success in 4-year-olds, whereas younger children succeeded with progressive alignment experience and/or linguistic scaffolding. They did not test children on the XY-CD problem, whereas full AMTS requires representations *both* of the relations same *and* different.

These findings on full 2-item AMTS, merely asserted by Premack, and on the relation same alone by Christie and Gentner, taken together, are consistent with the possibility that some form of representational change is necessary before children can succeed on the 2-item AMTS tasks. Perhaps there is a shift, between ages 3 and 6, from reliance on properties of the arrays such as entropy or

symmetry as a basis for solving AMTS tasks to reliance on representations of the relations among the elements. If this is so, 3-year-old children should resemble pigeons and baboons on AMTS tasks, failing at 2-item AMTS but succeeding at 16-item AMTS.

Human children learn linguistic symbols for the concepts *same* and *different* in the fourth year of life – i.e. the words "same" and "different" (Hochmann, Zhu, & Carey, unpublished data; Webb, Oliveri, & O'Keeffe, 1979). Perhaps learning the language to represent these relations in terms of the abstract summary symbols "same" and "different" plays an important role in the developmental changes reported by Premack. Christie and Gentner (2014) provide evidence for this possibility; priming 3- and 4-year-olds with linguistic labels for "same" and "different" improves performance on a subsequent XX-AA (vs BC) task. If so, this might be reflected in the explanations children provide for the basis of their choices in our AMTS tasks.

In sum, studies with children that are informed by the animal literature could provide a new wedge into the questions of what representations are involved in solving the 2-item and 16-item AMTS tasks, and whether the capacity to represent the abstract relations same and different in a format that support success in 2-item AMTS requires some sort of conceptual innovation, perhaps supported by language acquisition. The present studies begin such an exploration.

We explored preschoolers' ability to solve AMTS problems, comparing their performance to that of non-human animals and to that of human adults. The latter comparison requires a more complete picture of adult performance. Experiment 1 begins to fill the gap in our knowledge of adult performance on 16-item AMTS.

2. Experiment 1—Adults: 16-item array match to sample

For animals, rule discrimination tasks are in general easier than AMTS. For human adults, this might not be the case, for AMTS tasks invite comparison of the choice arrays (i.e., pairs of 16 same arrays and 16 different arrays) with respect to how they differ, and then invite analysis of the sample arrays in terms of a basis of matching it to one of the choice arrays that is consistent with how the choices arrays themselves differ. Engaging in this analysis would arrive at either the relations same and different among the elements or the property of entropy as a consistent basis of choice. In our version of the task, we of course provided no information as to whether the relevant difference between arrays should be entropy or the relations among the individual items. We seek to establish whether adults spontaneously solve 16-item AMTS tasks (perhaps even with no training, or within 8 training trials), even when 21% fail to learn 16-same vs. 16-different discriminations with up to 50 training trials (Castro & Wasserman, 2013), and if so, whether they recruit the relations same and different and/or entropy as a basis for matching.

2.1. Methods

2.1.1. Participants

We recruited participants online through Amazon Mechanical Turk. Ninety-nine participants (mean age = 39.32 years, SD = 12.57 years, 49 males) completed the study. One participant was excluded; this participant was a clear outlier, answering all 8 of the test trials incorrectly. Thus, there were 98 participants in the final sample. The large majority of participants identified as non-Hispanic Caucasian (80%), while the rest of the participants identified as black, Asian, or Hispanic.

2.1.2. Stimuli

We used arrays of 16 black-and-white symbols taken from a pool of 398 unique symbols, similar to those used by Fagot et al. (2001). The 16 items on each card were arranged in a four-by-four grid. There were five different types of arrays: *all-same (entropy 0)*, *4-different (entropy 1.3), 8-different (entropy 2.5), 12-different (entropy 3.5),* and *all-different (entropy 4)*.[2] In the *all-same* arrays, the 16 items were identical. In the *4-different* arrays, 12 of the items were identical, and 4 of the items were unique, and so forth for the *8-different* and *12-different* arrays. In the *all-different* arrays, all 16 items were unique (see Fig. 1).

The cards were grouped into triads of arrays enclosed in boxes, called "cards" in the instructions to the participants. A different triad was used on each trial. Within each triad, no individual item that appeared on one of the cards appeared on either of the other two. Two of the cards in each triad were designated *choice cards*. Choice cards always included one *all-same* and one *all-different* card. The third card in each triad was designated the *sample card*. During training, the sample cards were always either all-same or all-different arrays. During test the sample card could be any of the five array types (Fig.1). Participants were asked to select the choice card that they thought went with the sample card.

2.1.3. Procedure

2.1.3.1. Training trials. Participants received instructions on the first page of the study, as follows: "In this task, you will be shown some cards. Some cards go together and some cards don't go together. You will help sort them. In order to sort the cards correctly, you must look at the whole card. After you have selected a response, we will show you the correct answer." During the training trials, participants saw a screen showing sample card above and two choice cards below it, and were prompted to answer the question "Which of the cards below goes with the card above?" The participants then selected a choice card to match the sample card. Regardless of whether the participant's selection was correct or incorrect, the next screen showed the sample card and the correct

[2] The three disagreements all involved ambiguity between the "entropy" category and the "single item" category. One coder coded "multiple objects" in the entropy category; the other in the "single object" category. After discussion, this latter categorization was accepted, for "multiple objects" seems an implicit contrast to a single object. This choice was maintained throughout all 204 justifications.

choice card, with the feedback "Here's the answer: these two cards go together because these cards are more alike." Each participant went through a total of eight training trials, four with a all-same card sample and four with a all-different card sample. The procedure was counterbalanced such that the first training trial included an all-same sample card for half of participants and an all-different sample card for the other half.

Within each session the *all-same* choice cards appeared on the left side of the screen half the time and on the right side half of the time. In addition, the correct choice was 50% of the time the right card and 50% of the time the left card, such that a participant with a side bias would respond at chance both for all-same trials and for all-different trials. The order of all-same sample or all-different sample trials was randomized, subject to the constraint that there were no more than 3 trials of a single trial type in a row.

2.1.3.2. Test trials. The test trials were similar to the training trials, except no feedback was given. There were 20 test trials, 4 with each of the array types (*all same, 4-different, 8-different, 12-different,* and *all-different*) as samples (see Fig.1). The first two test trials consisted of an *all-same* sample and an *all-different* sample (to ensure that training was maintained when no feedback was given, before possible confusion due to intermediate arrays was introduced). Similarly, the last pair of test trials again included an *all-same* array and an *all-different* array as samples, to again ensure that training was maintained. We prompted participants for explanations for their choices for these last two test trials, asking "Why do you think the card you picked goes with the card above?" These justifications were asked independently of participants correct or incorrect choice.

2.2. Results

Adults overwhelmingly succeeded on the training trials, with an average of 99% correct on all-same trials and 98% correct on all-different trials. Given that virtually all participants (99%) made the correct choice on the first training trial, before any feedback, it is clear that adults spontaneously succeed at 16-item AMTS. Adults immediately saw that all-same (entropy 0) samples should be matched with all-same choices, and all-different (entropy 4) samples should be matched with all-different choices.

Analyzing the 8 test trials involving all-same or all-different samples as the training trials, 96 out of 98 participants (98%) chose the correct choice card on all 8. Another 2 participants (2%) scored 7 out of 8 correct. These results are in marked contrast with previous adult findings in the same-different discrimination literature, which show large proportions of adult samples failing at 16-item array discrimination tasks (16-same vs. 16-different) after 40 or 50 training trials (Castro & Wasserman, 2013; Castro et al., 2006). In subsequent studies with children, we adopt an 8/8 criterion for success on the test trials.

We next turned to the basis of adult matching, through an analysis of participants' choices when the sample cards were intermediate between 16-all-same samples and 16-all-different samples.

2.2.1. Intermediate sample trials

We divided participants into two groups based on a priori criterion those who matched the *all-same* sample cards with the *all-same* choice cards and all other sample cards with the *all-different* choice cards (*all-same/not all-same* sorters; the categorical responders in the studies from Young and Wasserman; Fig. 2b), and those who did not (alternative sorters). A significant proportion (22%) of the participants were categorical all-same/not all-same categorical responders. If we relax the criterion to allow one deviation from 100% adherence to this pattern, an additional 6 participants follow the all-same/not all-same rule (now 28 out of 98 participants (29%); data displayed on Fig. 5A). For all 6 of the additional participants, the deviation occurred at the 4-different array. This is the categorical pattern depicted in Fig. 2b, and while observed here, is markedly less common than in the same/different discrimination paradigm, where it is observed in 75% of 80% of participants who have first learned the 16-same/16-different discrimination.

The pattern of responding of the remaining 70 participants, or 71% of the sample, the "alternative sorters" is depicted on Fig. 5B. For these sorters, the proportion of selecting the all-different choice card was significantly different for all pairs of sample arrays (*Ps* $<$ .001, Bonferonni corrected) except for the 12-different and all-different sample arrays. However, this is not the logarithmic function depicted in Fig. 2a, and observed in both baboons and both humans in Fagot and Wasserman's 16-item AMTS task (Fig. 4). Notably, the difference between 4-different and 8-different trials was substantially larger than between all-same and 4-different trials, despite similar entropy differences of 1.2 and 1.3 respectively. Consequently, the alternative sorters are not using entropy as a

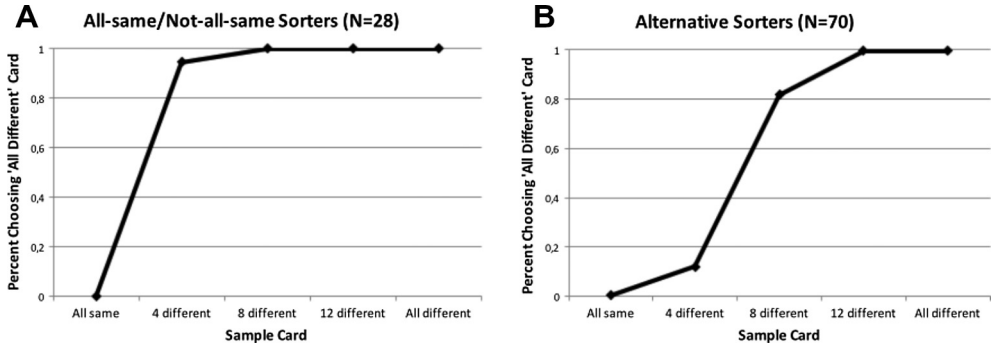

Fig. 5. Sorting patterns for all same/not all same sorters (A) and alternative sorters (B).

strategy for sorting intermediate arrays, as Weber's law dictates that the same difference of entropy should be more noticeable in a lower range. Weber's law thus predicts the largest difference between all-same and 4-different trials. Instead, the alternative sorters are most probably using the sorting strategy all-same vs. all-different, which does not yield a definite answer for intermediate cards. These must be sorted on the basis of each intermediate sample card's similarity to these two choice cards, which explains why 4-different is so close to all-same and 12-different is not significantly different from all-different, resulting in the pattern depicted on Fig. 2d). An analysis of participants' justifications for their last two choices (an all-same sample and an all-different sample) may shed further light on the basis of adults' choice.

2.2.2. Justifications

Participants were asked to explain their choice on the last two test trials, one with a 16-all-same sample card and one with a 16-all-different sample card. There were 204 justifications in total, because on 8 trials, a participant offered two explanations for their choice. Justifications were coded blind to what pattern of choices (categorical vs. alternative) the participant had made.

Justifications fell into 4 distinct categories: (1) *Irrelevant justifications* included non-responses ("I don't know" or no justification provided) and irrelevant justifications, in which the participant merely restated their judgment, e.g., "they go together because they are more alike" or didn't even say something with the form of an explanation, such as "yes." These were rare (7% of all justifications). (2) *Entropy justifications* (21% of total) mentioned that the chosen card matched the target on the basis of being or not being a mixture of objects, random, variable etc., with no mention of the relations "same" or "different." (3) *Single object justifications* (10% of total) contrasted the two choices on the basis that one of them contained a single icon, only one icon, or named the single icon (e.g. "all hats") whereas the other contained multiple objects. (4) *"Same/different" language justifications* (63% of total). These latter justifications varied in the degree of explicit elaboration that the relations sameness and difference among icons on each card are the basis of the match. Almost all used the language of "same" and "different" (e.g., "all of the icons are the same as each other" to "they are both all different" to "the icons are the same" or "the icons are different.") Justifications that used the language "match," "are alike," "repeat" as in "all matching symbols" or "none of the icons are alike," or "a single item is repeated throughout" were also coded in this category. One coder (SC) coded all 204 justifications; another coder (RZ) coded 50 of them. Intercoder agreement was 94%[2].

A notable feature of the justifications was a very high level of explicit quantificational language in the single object and same/different language categories. Sixty percent of the single object justifications and 65% of the same/different language justifications included explicit quantifiers that indicated quantifying over individuals in a set (e.g., single object justifications: "only one type of symbol," "all one symbol;" same/different language justifications: "all objects in the box are the same", "none of the images are similar.") In contrast, quantificational language was used rarely in entropy justifications (10%, e.g., "all mixed"). This shows that "same" and "different" express relations among individuals within sets, and suggests that single object justifications expressed the same content as did same/different language justifications.

In sum, by far the dominant justification type (63%) included explicit language of relations, such as "same", "different," "similar," "alike" and "match." Nonetheless, although much rarer (21% overall), participants also articulated that the icons differed in entropy in their justifications. If alternative sorters (Fig. 5B) are matching on the basis of entropy, whereas categorical all-same/not all-same sorters (Fig. 5A) are sorting on the basis of the relations same and different, one would expect that justifications for each pattern would differ. The proportion of irrelevant and entropy justifications were virtually identical across the two patterns of sorting, (irrelevant: 5%, 6%; entropy: 21%, 20%, for categorical and alternative sorters, respectively). Categorical all-same/not all same sorters were more likely to provide single object justifications than were alternative sorters (20% vs. 6%) and slightly less likely to provide "same/different" language justifications (54% vs. 68%). These differences were significant; $\chi^2(1) = 7.36$; $P < .01$. Indeed one-object justifications fit particularly well an all-same vs. not-all-same categorical rule.

These analyses indicate that participants' explanations do not straightforwardly reflect the basis of judgment—participants do not articulate a rule induced during training and followed during test. Rather, participants are articulating a difference between the two choice cards that is consistent with the choice they made. The important results are that by far the dominant justifications are articulated in terms of explicit references to the relations same and different, *both* for alternative and categorical all-same/not all-same categorical sorters. Just as the details of the pattern of choices alternative sorters make to intermediate arrays (Fig. 5B) suggest matches are being made on the basis of the contrast between all-same and all-different, rather than entropy, the justifications provide no evidence that entropy representations are driving responses of alternative sorters.

The 16-item AMTS task of Experiment 1 yields strikingly different performance from the same-different array discrimination tasks previously studied with human adults (Castro & Wasserman, 2013; Castro et al., 2006). Firstly, adults in the AMTS task showed spontaneous success, sorting all-same with all-same and all-different with all-different with near-perfect accuracy on the very first trial. In stark contrast, large proportions of adult samples fail to learn the rules such as "if all same, press left key; if all different press right key" at 70% or 75% accuracy even after 40–50 trials with feedback as to the correct choice (Castro & Wasserman, 2013; Castro et al., 2006). We assume that the AMTS task contrasts with the same-different discrimination task because adults approach it with the strategy of a double comparison – choices with each other, each choice with sample – constrained by a search for a single dimension of difference between the choices, such that the sample matches one of the choices on that dimension.

Secondly, whereas around 75% of the participants in the rule discrimination tasks who learn a rule based on the contrast of 16-same vs. 16-different, sort intermediate arrays according to the rule: "all same/not all same", in Experiment 1, only around 30% displayed this pattern of sorting. Those who did not, the "alternative sorters", used another rule on the basis of abstract relations, specifically "all same/all different". In sum, Experiment 1 shows that adults easily learn to match arrays on the basis of same and different and do not use entropy as a sorting strategy. Thus, these results also contrast with the only published previous study on 16-item AMTS with adults. Unlike the 2 adults in Fagot, Young and Wasserman (2001), who appeared to choose based on entropy, we

found no evidence, among the almost one hundred participants in the present study, of a reliance on entropy as a basis of choice. Why the difference? Again, it is wrong to think of performance on these tasks as reflecting learning a rule during training, and then applying it during test—the adult search for a single consistent basis of choice that applies to all the stimuli, including the new intermediate arrays, continues during the test. Fagot, Young and Wasserman (2001) included many more intermediate levels of entropy, and very different methods of producing arrays with a given entropy. This most likely makes entropy a more viable hypothesis concerning the relevant dimension of variation. Human adults clearly *can* represent the entropy (degree of variability) within an array, and can use it as a basis for choice in AMTS tasks.

In sum, Experiment 1 gives us what we want in our exploration of pre-schoolers' performance in AMTS: a 16-item AMTS that is spontaneously solved by human adults, and which happens to be, but not logically necessarily, solved on the basis of representations of the relations same and different by these adults. Experiments 2–5 now turn to comparisons of pre-schoolers with non-human animals and with human adults on AMTS.

3. Experiment 2 – 2-item array match to sample

Non-human animals perform *much* better on 16-item (all-same vs. all-different) AMTS than they do on 2-item (both same vs. both different) AMTS. Confirming the generalization reported by Premack (1983, data not published), in a task that could also be approached as a same-different discrimination task in which children must learn to "pick same", Christie and Gentner (2014) showed that 3-year-old do not spontaneously succeed at a partial 2-item AMTS (same targets only). Experiment S1, reported in Appendix A, found that 3- and 4-year-olds failed at a full 2-item AMTS task, one which also could be approached as a conditional same-different discrimination task, learning the rules "doggie likes same" and "bear likes different." Five-and 6-year-olds, in contrast, succeeded. The results of Experiment S1 are congruent with the pattern reported by Premack (1983): 3- and 4-year-olds failed, while 5- and 6-year-olds succeeded. Experiment 2 explores the behavioral change between 4 and 5 years on 2-item AMTS, when the *only* basis of choice is matching to sample. Experiment 3 then explores whether for children, like for non-human animals, performance is better on 16-item AMTS, and whether the basis of success is representations of the relations same and different or representations of entropy.

3.1. Methods

3.1.1. Participants

Children who participated in Experiments 2–5 were identified from publically available birth records and families were invited to bring their children to the laboratory. These participants were over 65% non-Hispanic Caucasian, with the rest self-reporting as black, Hispanic, Asian, or mixed. A total of 24 4-year-olds (M = 4.51 years; SD = 0.25; range = 4.02–4.96 years; 13 girls) and 24 5-year-olds (M = 5.53 years; SD = 0.30; range = 5.00–5.98 years; 10 girls) participated in Experiment 2. Researchers tested two additional 4-year-olds, whose data was excluded due to experimenter error.

3.1.2. Stimuli

We used two types of arrays, printed on small laminated flashcards: same-arrays, in which the two icons were identical, and different-arrays, in which the two icons were non-identical. Each card was used on only one trial, and no icon appeared on more than one card. To minimize the perceptual confound of symmetry, the icons were randomly placed vertically or diagonally (see Fig. 3 for examples).

3.1.3. Procedure

3.1.3.1. Training trials. The experimenter introduced the study by saying, "We're going to play a game. I'm going to show you some cards. Some cards go together and some cards don't go together. I'm going to ask you to help me sort them. First I'll teach you how to play the game."

The experimenter then held up a same-array choice card and said, "See this card?" and placed it on the table. After the child responded affirmatively, the experimenter held up a different-array choice card and said, "See this card?" and also placed it on the table. While touching both choice cards, she said: "These cards do not go together because these cards are not alike." Then, the experimenter held up a third, sample, card, either with a same- or different-array, and placed it on the table, saying "See this card?" Pointing back at the two choice cards, the experimenter said, "Which one of these cards goes with this [the sample] card?"

If the child made a correct response, the experimenter said, "That's right! These cards go together because these cards are more alike" and moved onto the next trial. If the child made an incorrect response, the experimenter said, "Nice try, but actually these cards don't go together because these cards are not alike." The experimenter then took away the incorrect card and showed the child the two cards that matched, stating, "These cards go together because these cards are alike. In this game, you have to look at the whole card to figure out which ones go together. Do you see why these cards go together?"

Each child went through a total of eight training trials, four with a same-card sample and four with a different-card sample. The procedure was counterbalanced exactly as the training trials were in Experiment 1.

3.1.3.2. Test trials. The test trials were identical to the training trials, except the experimenter stopped providing feedback. The experimenter introduced this phase by saying, "Alright, now that you know how to play the game, I'm going to let you play all by yourself! That means I'm not going to tell you which cards I think go together anymore and just let you choose." The experimenter then proceeded through the test trials, but did not correct the child's mistakes. Rather, neutral, positive feedback was used

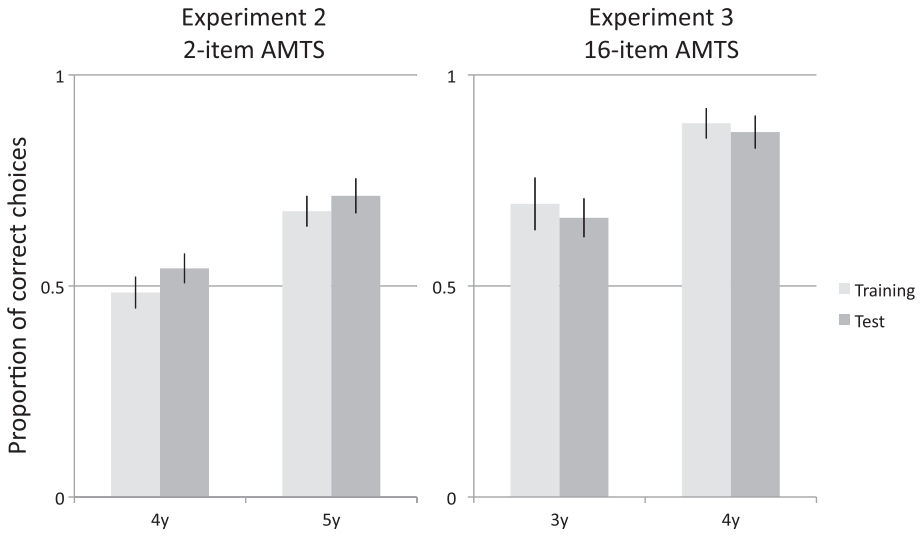

Fig. 6. Results of the training and test in Experiments 2–3. For Experiment 3, only the results of the test on all-same/all-different cards are presented. Error bars represent standard errors from the means.

throughout (e.g., "you are doing a great job." "Good work."). Each child received eight test trials, four with a same-sample card and four with a different-sample card, counterbalanced as were the training trials. For the last two test trials, one with a same-sample card and one with a different-sample card, the child was asked for an explanation of his or her choice ("Why do you think this card goes with this card?"), independent of the choice being correct or incorrect.

3.2. Results

The results of Experiment 2, shown in Fig.6, confirm the results obtained in Experiment S1 (Appendix A). Four-year-olds failed the 2-item AMTS task, while 5-year-olds succeeded.

A repeated-measure ANOVA with Trial Type (Same, Different) and Experimental Phase (Training, Test) as within-subject factors and Age (4 years, 5 years) as between-subject factor revealed a main effect of Age; $F(1, 46) = 13.47$; $P = .001$. Other main effects and interactions were not significant ($Ps > .15$). Furthermore, 5-year-olds performed better than chance on test trials, collapsing across same-trials and different-trials (as there were no effects of Trial Type): 71% correct, $t(23) = 4.46$; $P < .001$ corrected for 2 comparisons. In contrast, 4-year-olds' performance did not differ from chance on test trials: 54% correct; $t(23) = 0.98$; $P = .66$ corrected for 2 comparisons.

The ANOVA above found no effects of experimental phase; meaning that performance was the same on the training trials as on the test trials. We confirmed that 5-year-olds were also above chance on the training trials (68% correct, $t(23) = 4.38$, $p < .001$, corrected for 2 comparisons) whereas 4-year-olds were at chance (48% correct, $t(23) = 0.44$, $p = 66$, uncorrected). Thus, success in this task, for those who achieved it, was almost immediate, leading to statistically equivalent performance over the 8 training trials as observed on the 8 test trials.

Experiment S1 and Experiment 2 were quite different from each other in the way AMTS was probed, in the number of trials, and in the nature of and amount of feedback, and yet the two studies converged on the finding that 5-year-olds, as a group, succeeded at 2-item AMTS (but were far from being at ceiling, 67% correct in Experiment S1; 71% in Experiment 2), whereas children age 4 or younger did not (age 4: 54% correct in both studies).

Our final analyses sought to more fully understand the source of the improvement from age 4 to 5. We divided children into "succeeders" who got 100% (8/8) of the test trials correct ($P < .01$, binomial test 2-tailed) and "non-succeeders" who did not (Table 1). There were 7 succeeders among the 24 5-year-olds (a number significantly more than one would expect by chance; $P < .001$, binomial), compared to only 1 among the 4-year-olds (a number that might well arise by chance, $p = .17$, binomial). The overall level of performance among *non-succeeders* did not differ from chance; 5-year-olds: $n = 17$, 60% correct on test trials, $t(16) = 2.34$; $P = .07$ corrected for 2 comparisons; 4-year-olds: $n = 23$, 52% correct on test trials, $t(22) = 0.56$; $P = .58$ uncorrected. Thus, children either understood a relevant basis of choice on this task, succeeding at all of the test trials, or they did not, and performed at chance. That succeeders were qualitatively different from non-succeeders was confirmed in an analysis of children's justifications for their choices.

Asked why they sorted as they did, children gave explanations codable into five distinct categories, four of which were also observed in the adult explanations in Experiment 1. Shared with adults, were Uninformative justifications: Don't know, No response, Uninterpretable (e.g., "Because he kissed her)," or mere Restatements of the judgment, (e.g., "because they are alike," "because they match," "because they are the same," "because they go together," "or because they look alike)." Such justifications were uninformative as to *any* possible basis for the choice. Also, like adults' explanations for their choices, there were three types of

Table 1
Number of Succeeders (100% correct) in each Experiment; Percent Correct by Non-Succeeders in each Experiment.

	Number of succeeders (of 24) (100% correct)			% Correct by non-succeeders		
	Age 3	Age 4	Age 5	Age 3	Age 4	Age 5
Exp 2 2-item	–	1	7	–	52%^{ns}	60%^{ns}
Exp 3 16-item entropy 0/4	2	13	–	63%*	70%**	–
Exp 4 2-item	0	5	10	54%^{ns}	59%^{ns}	54%^{ns}
Exp 5 16-item entropy 0/1	1	7	–	50%^{ns}	51%^{ns}	–
Exp 5 2-item	1	4	–	47%^{ns}	60%^{ns}	–

** $P < .01$.
* $P < .02$.
^{ns} $P > .05$.

justifications relevant to the correct choices. Of these relevant justifications, the most common category was Same/Different Language justifications: children explicitly mentioned that icons *within each card* were "the same" or "different" (e.g., "because the two on this card are the same and the two on this card are the same" or "because on both of them the pictures are different."). Children also provided One Object justifications: justifying a correct same-target, same-choice card match, by comments such as "because both of these are clocks and both of these are hammers;" ("All/both X" or "Not all/both Y"). Finally, across Experiments 2–5, there were two explicit references to entropy (e.g., "because this one is random and this one is random"; designated "Entropy"), although none of these occurred in Experiment 2. The fifth category, ("Features") was not observed in the adult justifications in Experiment 1, and reflected attention to matches at the level of features of individual objects (e.g., "because this one is round and this one is round," indicating single images on two different cards) or features of the whole arrays unrelated to entropy or the relations same and different (e.g., "because this one is black and this one is black," indicating the whole cards, where in fact, all three cards were black and white).

For Experiments 2–5, two independent coders labelled each of the children's justifications as one of the 5 categories described above (Uniformative, Feature, "Same/Different" Language, One Object, Entropy). The two coders reached 90% intercoder agreement. Disagreements were resolved by review and discussion.

We grouped together Uninformative and Feature justifications, into a category of Irrelevant justifications as neither invokes information relevant to *success* at the task. We also grouped "Same/Different" Language and "One Object" Justifications together into a category of Relational justifications, because these patterned together in the adult justifications, and children, like adults, frequently used quantifiers in each type of justification, which indicates quantifying over individuals in the arrays ("both the same", "both hats.") The first result of note is children's justifications were vastly more likely to fall in the Irrelevant category (4-year-olds, 97%; 5-year-olds, 70%) than were adults' justifications (7%). This is consistent with the fact that children overwhelmingly failed to match on the basis of either the relations sameness/difference or entropy.

In Experiment 2 (and in all subsequent experiments), we divided the children into "succeeders" and "non-succeeders", collapsing across age, as non-succeeders were at chance on 2-item AMTS regardless of age, and we analyzed justifications to assess the bases of each group's response. Here we compare the proportion of participants who gave justifications that appealed to Relational (explicitly in "Same Different" Language justifications or implicitly in One Object justifications) versus Irrelevant justifications (Uniformative and Feature justifications.) As can be seen on Fig. 7, in Experiment 2, virtually all of the succeeders gave a Relational justification, and the overwhelming majority of these were explicit appeals to the relations same and different, using the words "same" and "different." In contrast, *every one* of the non-succeeders' justifications was categorized as Irrelevant. That is, at both ages, non-succeeders, who were at chance matching same to same and different to different, gave uninformative justifications, or referred to features of individuals or of the arrays (see Appendix B for a full breakdown of the justifications in all experiments by age, and by succeeder/non-succeeder status).

In sum, the overall better performance of 5-year-olds as a group than 4-year-olds is *entirely* due to a small group of children (n = 7), who were perfect on the test trials and articulated the rule they were following, explicitly stating that they were matching on the basis of the concepts *same* and *different*.

4. Experiment 3 – 16-item array match to sample

Like young children, baboons and pigeons generally fail the array match-to-sample with 2-item arrays (2-item AMTS); however, when the task employs 16-item arrays, baboons and pigeons successfully match all-same and all-different sample arrays to the respective choice arrays (16-item AMTS; pigeons: Cook & Wasserman, 2007; baboons: Fagot et al., 2001). In addition, when trained with 16-same and 16-different sample and choice arrays, and tested with sample arrays comprising mixtures of identical items and

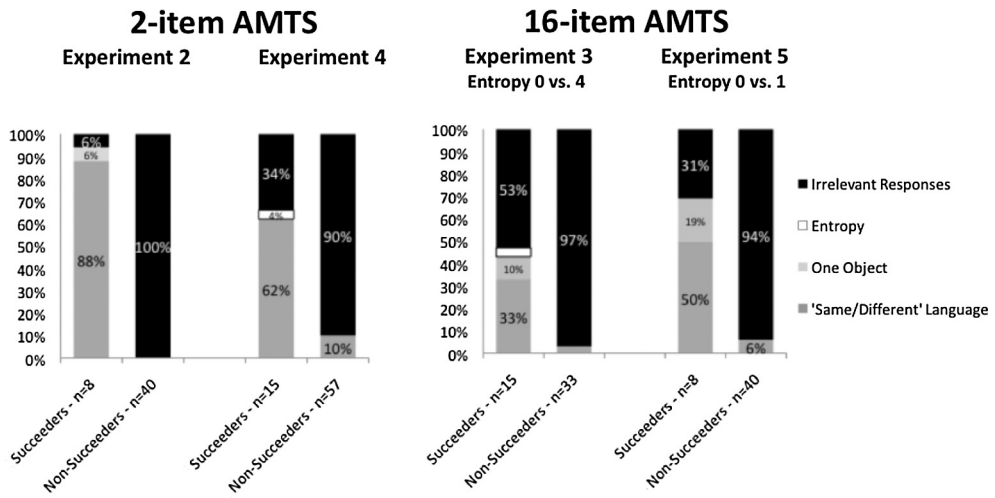

Fig. 7. Distribution of the different types of justifications, collapsed across ages, for the last two 2-item AMTS tests in Experiments 2 and 4 and for the last two 16-item AMTS tests in Experiment 3 (all-same vs. all-different) and Experiment 5 (all-same vs. 13-same/3-different).

different items, baboons do not induce the all-same vs. not all-same categorical rule, but instead match on the basis of entropy (pigeons have not been tested in this version of the task).

Experiment 3 asks whether 3- and 4-year-olds, who failed in 2-item AMTS (Experiments S1 and 2), succeed in essentially the same 16-item AMTS task as that of Experiment 1, and like adults do so with little or no training. If these young children successfully match 16-item all-same and all-different arrays, it would suggest that either the larger arrays of all-identical or all-different icons make the relations same/different more salient, or that young children, like non-human animals, are able to match based on representations of entropy. To distinguish between these possibilities, we analyze both children's pattern of performance on 16-item AMTS when the sample cards include mixtures of identical items and different items and their justifications for their choices.

4.1. Methods

4.1.1. Participants

A total of 24 3-year-olds (M = 3.47 years; SD = 0.31; range = 2.95–3.99 years; 9 girls) and 24 4-year-olds (M = 4.54 years; SD = 0.30; range = 4.00–4.99 years; 16 girls) were drawn from the same participant pool as in Experiment 2. Three additional 3-year-olds were tested but failed to complete the experiment.

4.1.2. Stimuli and procedure

The stimuli and procedure, including counterbalancing, were essentially identical to those of Experiment 1, with the following modification. The arrays were displayed on cards made of laminated pieces of 8.5″ × 11″ paper, and the instructions and training were modified for young preschoolers, such that they were identical to those of Experiment 2. The eight training trials included four 16-same sample cards and four 16-different sample cards. There were 20 test trials, 4 with each of the array types on Fig. 1 as samples. The counterbalancing and ordering of test trials was the same as in Experiment 1. Specifically, the first two test trials consisted of an all-same sample and an all-different sample, as did the last two test trials, and children's explanations for their choices were elicited for these last two test trials. The order of the remaining test arrays, presented between the first pair and last pair of test trials was separately randomized for each child.

4.2. Results

The results were analyzed in three steps. First, we asked whether children learned the 16-item AMTS rule, looking both at training trials and at the 8 test-trials where the sample trials were either all-same (n = 4) or all-different (n = 4). Unlike adults, there were many non-learners (non-succeeders, who fail to get 8 of 8 test choices correct). However, like pigeons and baboons, and unlike the 4-year-olds in Experiment 2 (2-item AMTS) and 3- and 4-year-olds in Experiment S1, both 3-year-olds and 4-year-olds were above chance on 16-item AMTS. Furthermore, children learned quickly enough to be above chance even on the eight training trials. Second, we analyzed the pattern of responses on the intermediate mixed arrays, and third we analyzed children's justification for their last two choices (one all-same sample and one all-different sample). With respect to intermediate arrays and justifications, there is evidence that some children succeeded on the basis of representations of same and different (unlike animals, and like adults), and some on the basis of entropy (like animals, and unlike adults). Detailed analyses follow.

4.2.1. Learning 16-item AMTS

Fig. 6 displays the percent correct on the 8 training trials and the 8 test trials with all-same or all-different samples. A repeated-measure ANOVA examined the effects of Phase (Training, Test), Trial Type (all-same sample, all-different sample) and Age (3 years; 4 years) on percent correct on 16-item AMTS. It revealed a main effect of Age; $F(1, 46) = 11.87$; $p = .001$, with 4-year-olds ($M = 86\%$) performing better than 3-year-olds ($M = 68\%$). Other main effects and interactions were not significant ($P > .05$).

As always, we seek not only to establish developmental improvement, but also to learn the age of success. Three- and 4-year-olds, as groups, performed better than chance on test trials. Collapsing across all-same and all-different trials (as there were no effect of Trial Type), both 3- and 4-year-olds performed better than chance (respectively 66% correct; $t(23) = 3.51$; $P < .01$ corrected for 2 comparisons; and 86% correct; $t(23) = 9.34$; $P < .0001$ corrected for 2 comparisons).

The ANOVA above found no effect of experimental phase; meaning that performance was the same on the training trials as on the test trials. We confirmed that the 3- and 4-year-olds were each above chance on the training trials (respectively 69% correct; $t(23) = 3.11$; $P = .01$ corrected for 2 comparisons; and 89% correct; $t(23) = 10.70$; $P < .0001$ corrected for 2 comparisons). Thus, like pigeons and baboons, 3- and 4-year-olds' performance on 16-item AMTS is much better than that on 2-item AMTS.

We next explored whether, as is the case for pigeons and baboons, this success relies entirely on entropy, or like the human adults in Experiment 1, it is entirely due to representations of the relations same and different. First, we divided children into *succeeders* (8/8 test trials correct, 100%) and *non-succeeders* (everybody else). As can be seen in Table 1, there were vastly more 4-year-old *succeeders* on 16-item than on 2-item AMTS (13 vs. 1, $P < .001$, Fisher's exact test, 2-tailed), confirming that 16-item AMTS is much easier for young preschoolers than is 2-item AMTS. There were even two 3-year-old succeeders. Also, whereas both 4- and 5-year-old non-succeeders were *at chance* on 2-item AMTS in Experiment 2, as can be seen in Table 1, both 3-year-old and 4-year-old non-succeeders performed *better than* chance on 16-item AMTS (3-year-olds: $n = 22$, 63% correct on test trials, $t(21) = 2.93$; $P = .016$ corrected for 2 comparisons; 4-year-olds: $n = 11$, 70% correct on test trials, $t(10) = 3.79$; $P < .01$ corrected for 2 comparisons). That is, even 3- and 4-year-old *non-succeeders* draw on some relevant representations of 16-all-same (entropy 0) and 16-all-different (entropy 4) sample cards to match them to the correct choice cards more than would be expected by chance.

Analyses of justifications also revealed a subtly different pattern from those of Experiment 2 (Fig. 7; see Appendix B for the full analysis of children's justifications). Again, we group Uninformative and Feature justifications together as Irrelevant. All but one of the 3-year-olds' responses were Irrelevant. As in Experiment 2, succeeders in Experiment 3 were more likely to give a justification on the basis of relations than were non-succeeders, but the difference was not as categorical. In Experiment 3, 4-year-old succeeders were five times as likely to give Relational justifications (42% overall; "Same/Different Language: 35%, Implicit Relational: 8%) than were *non*-succeeders (9%), but not all succeeders articulated a basis for their match in terms of the relations same and different. In Experiment 2, in contrast, the ratio of Relational justifications by succeeders to non-succeeders was 94% to 0%. It is possible, then, that while some children induced an explicit rule based on the relations same and different to successfully match target arrays with the all-same and all-different 16-item choice arrays, others may have relied on alternative representations (e.g., entropy) to do so. We now turn to the analyses of the intermediate entropy sample cards to further explore this possibility.

4.2.2. Performance on 4-different, 8-different, and 12-different test trials

In Experiment 1, the patterns of responding across all five types of samples suggest that all adults matched the sample to the choices on the basis of the relations same and different (29% all-same vs. not all-same pattern, pattern of Fig. 2b; 71% all-same vs. all-different pattern, pattern of Fig. 2d). We first explored whether any children matched samples to choices according to the distinction between all-same vs. not all-same samples. We credited children with this pattern according to the following conservative criterion: 100% choice of all-same array when probed with an all-same sample, and at most 1 deviation from 100% choice of all-different array for all other arrays ($P < .001$ binomial test, 2-tailed). Seven of the 4-year-olds (29%; the same proportion as adults) met this criterion; none of the 3-year-olds did. However, one 3-year-old had only 2 deviations from it ($P < .001$ binomial test, 2-tailed), and no other child at either age did so, so we also placed this child in this category. All remaining patterns were classified as "alternative patterns" in which intermediate samples were often sorted with the all-same choice.

By the selection criterion for categorical all-same/not all-same responders, these were all succeeders (sorting all 16-same test samples with the 16-same choice and all 16-different test samples with the 16-different choice). But not all succeeders provided this

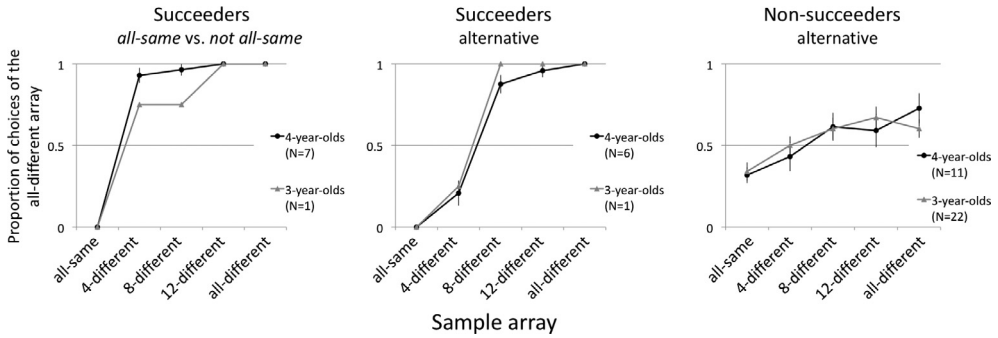

Fig. 8. Results of the test in Experiment 3: proportion of *all-different* choices for sample of various entropies. Error bars represent standard errors from the means.

pattern of choice; half provided alternative patterns. Furthermore, *all* non-succeeders provided alternative patterns. But as inspection of Fig. 8 reveals, for both 3- and 4-year-olds, the alternative patterns of non-succeeders differ qualitatively from those of the succeeders. An ANOVA involving only those children with *alternative patterns* examined the effects of Target Type (all-same, 4-different, 8-different, 12-different, and all-different) and Succeeder Status (succeeder – n = 7 with alternative patterns – vs. non-succeeder – n = 33) on the percent choice of the all-different choice card. There was a main effect of Target Type, as the percent choice of the all different-sample increased with increasing entropy of the sample; $F(4, 152) = 39.35$; $P < .001$. Importantly, the interaction between Target Type and Succeeder Status was also significant; $F(4, 152) = 12.60$; $P < .001$. Part of the interaction was due to significant differences between the two groups at correct classification of the 16-same and 16-different samples (i.e., was due to succeeder status itself; *Ps* < .002). The two groups also differed significantly on the 4-different, 8-different and 12-different samples (*Ps* < .02). For all-same and 4-different sample cards, Non-succeeders made more all-different matches than did Succeeders with alternative patterns. Conversely, for 8-different, 12-different and all-different targets, Non-succeeders made fewer all-different matches than did Succeeders with alternative patterns. As is evident from Fig. 8, The alternative pattern of the Succeeders closely matches the alternative pattern of adults in Experiment 1, and is consistent with following a matching rule in which all-same samples go with all-same choices and all-different samples go with all different choices, with intermediate samples (for which this rule does not apply) sorted by similarity to the two choices. As with adults, categorical all-same/not all same sorters were more likely to provide single object justifications than were alternative sorters (19% vs. 0%) and less likely to provide "same/different" language justifications (25% vs. 35%).

One last exploratory analysis further explored whether the signal that leads to above chance performance by non-succeeders (all of whom provide alternative patterns) is due to entropy. These data are too sparse to attempt to statistically examine whether the *best* fit of the function from entropy of the target to percent match to the all-different choice card is logarithmic. However, exploratory analyses fit linear (Match to all-different = a + b ∗ entropy) and logarithmic (Match to all-different = a + b ∗ log(entropy + 1)) functions to the alternative patterns of both 3- and 4-year-old non-succeeders. Both functions provided good fits, but the logarithmic function was slightly better ($R^2 = 0.98$; root mean squared error = 0.020) than the linear function ($R^2 = 0.94$, root mean squared error = 0.038).

In sum, the results of Experiment 3 testing children on 16-item AMTS contrast with the results on 2-item AMTS in Experiment 2. Most importantly, both 3- and 4-year-olds succeed statistically in Experiment 3, in the face of abject failure from 4-year-olds in Experiment 2 (and both 3- and 4-year-olds in Experiment S1). Moreover, the statistical success in Experiment 3 is due to the contribution from two quite different populations within our 3- and 4-year-olds. A small proportion of these children, like the adults in Experiment 1, demonstrated spontaneous, immediate, and total success. These are our succeeders—100% correct on the test trials with 16-same and 16-different samples, and near ceiling performance on the training trials as well. These children, both 3- and 4-year-olds, exclusively relied on representations of same and different—sorting matching on the basis of a categorical all-same vs. not all-same rule, or an all-same vs. all-different rule. Non-succeeders performed relatively poorly when presented with all-same or all-different 16-item samples, and provided a high proportion of irrelevant justifications. But unlike in Experiment 2, non-succeeders (33 of the 48 3- and 4-year-olds) still performed better than chance, and their pattern of responses across the intermediate sample cards was consistent with entropy being the source of their statistical success.

5. Experiment 4 – Transfer from 16-item arrays to 2-item arrays

To further explore the representations behind 16-item AMTS success in preschool aged children, in Experiment 4, we tested 3-, 4- and 5-year-olds on 2-item AMTS after training them on 16-item AMTS. Rules based on all-same and all-different apply both to all-same/all-different 16-item arrays and to same/different 2-item arrays. In contrast, the entropy of all-different 16-item and 2-item arrays differ (4 vs. 1). Failure to generalize successful sorting to 2-item AMTS, especially for *different* cards, would suggest that children were using entropy as the basis of their successful categorization in the 16-item AMTS task, rather than any rule formulated over representations of same and different.

When trained to match 16-all-same sample cards to 16-all-same choice cards and 16-all-different cards to 16-all-different choice cards, the results of Experiment 3 suggest that about half of the 4-year-olds and only two 3-year-olds learned a rule relying on the relations same and different: either all-same vs. not all-same, or all-same vs. all-different. Both of these rules should lead to successful generalization to 2-item AMTS, and should be reflected in explicit justifications formulated over the relations same and different.

In contrast, the pattern of response of non-succeeders in Experiment 3 (almost all 3-year-olds and half of the 4-year-olds) was consistent with entropy being responsible for their above chance performance on the test trials. For those children, training on all-same vs. all-different 16-item AMTS should not transfer to 2-item AMTS, as is found in the animal literature (Wasserman & Young, 2010).

5.1. Methods

5.1.1. Participants

A total of 24 3-year-olds (*M* = 3.50 years; *SD* = 0.30; range = 3.02–3.93 years; 14 girls), 24 4-year-olds (*M* = 4.55 years; *SD* = 0.30; range = 4.05–4.96 years; 11 girls), and 24 5-year-olds (M = 5.57 years; SD = 0.25; range = 5.00–5.97; 8 girls) participated in this experiment. Over 68% of the participants were non-Hispanic Caucasian, with the rest self-reporting as Hispanic, Asian, or mixed. We tested four additional 3-year-olds, who were excluded from the sample due to fussiness (one child), experimenter error (one child), or an inability to understand the instructions (two children).

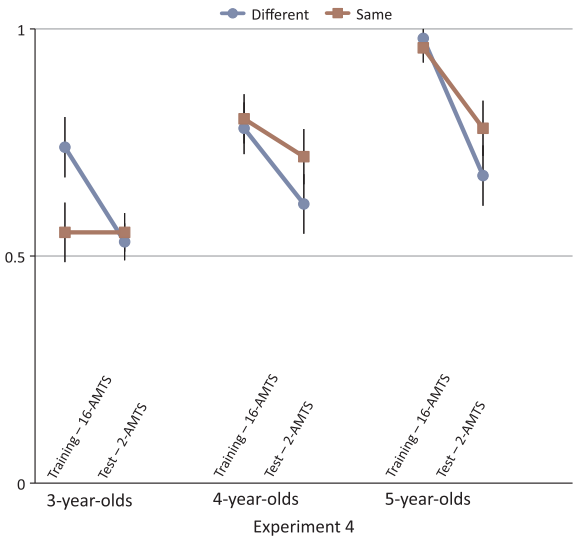

Fig. 9. Results of the training and test in Experiments 4. Error bars represent standard errors from the means.

5.1.2. Stimuli and procedure

The procedure was the same as for Experiment 2, except that the experimenter used the 16-item all-same and all-different arrays from Experiment 3 instead of 2-item arrays during the training trials. Test trials were identical to those of Experiment 2.

5.2. Results

Fig. 9 displays the results of Experiment 4. We compared performance on 16-item AMTS during training with performance on 2-item AMTS in test, to explore whether the basis of success on the two tasks differs. These analyses confirm that 16-item AMTS is easier than 2-item AMTS. Three-, 4- and 5-year-olds all succeed at 16-item AMTS within the 8 training trials. Success on subsequent 2-item AMTS test-trials was less strong: absent altogether at age 3, robust for 5-year-olds, and observed for the first time in this series of experiments among 4-year-olds. Detailed analyses follow.

A repeated measure ANOVA examined the effects of Experimental Phase (Training with 16-item arrays vs. Test with 2-item arrays), Trial Type (Same, Different), and age (3 years, 4 years, or 5 years) on the dependent measure of proportion of correct responses. There was a main effect of Age; $F(2, 69) = 14.484$; $P < .001$. Bonferroni post hoc tests showed that 5-year-olds performed better than 4-year-olds ($P = .042$) and 3-year-olds ($P < .001$), and 4-year-olds performed better than 3-year-olds ($P = .017$). We also observed a main effect of Experimental Phase, with children performing better in the Training phase with 16-item arrays than in the Test phase with 2-item arrays; $F(1, 69) = 21.027$; $P < .001$. Finally, there was an interaction of Experimental Phase and Trial Type; $F(1, 69) = 8.487$; $P = .005$, due to children performing better on different-trials in the Training phase with 16-item arrays than in the Test phase with 2-item arrays (Training = 83% correct; Test = 61% correct; $t(71) = 5.10$; $P < .001$ Bonferroni-corrected), while the performance on same-trials in the Training phase with 16-items and in the Test phase with 2-items did not differ significantly (Training = 77% correct; Test = 68% correct; $t(71) = 2.23$; $P = .17$ Bonferroni-corrected). As noted above, 16-item and 2-item all-same cards do not differ in entropy, whereas 16-item and 2-item all-different cards do differ in entropy (4 vs. 1). Thus, the observed interaction of Experimental Phase and Trial Type is compatible with some children relying on entropy. There was no other significant main effect or interaction (all $Ps > .07$). Importantly, the lack of interaction between Age and Experimental Phase ($F(2, 69) = 1.53$; $P = .22$) showed that children performed better on 16-item trials than on 2-item trials equally throughout the tested age range, from 3 to 5 years of age.

As always, we are not mainly interested in whether children improve with age, but rather in age of success. On the training trials on 16-item AMTS, 3-, 4- and 5-year-olds performed better than chance (3-year-olds: 65% correct, $t(23) = 2.70$; $P = .04$ corrected for 3 comparisons; 4-year-olds: 79% correct, $t(23) = 6.08$, $P < .001$, corrected for 3 comparisons; 5-year-olds: 97% correct, $t(23) = 17.80$; $P < .001$ corrected for 3 comparisons).

On the test trials on 2-item AMTS, 3-year-olds performed at chance (54% correct, $t(23) = 1.88$; $P = .22$ corrected of 3 comparisons), while 4- and 5-year-olds performed better than chance (4-year-olds: 67% correct, $t(23) = 3.52$, $P < .01$, corrected for 3 comparisons; 5-year-olds: 73%, $t(23) = 3.98$; $P < .01$ corrected for 3 comparisons).

As in previous experiments, we asked whether success on 2-item AMTS depends entirely on a small group of children who induced an explicit rule formulated in terms of the relations same and different. Children were asked for justifications only on the last two test trials (2-item AMTS). There were no 3-year-old succeeders (100% correct in the 8 2-item AMTS test trials). Five 4-year-olds and 10 5-year-olds were succeeders, a greater number than in Experiment 2 (see Table 1, and see below). Table 1 also shows that, as in Experiment 2, non-succeeders on 2-item AMTS at every age were at chance on the test trials (3-year-olds, n = 24, 54% correct, $t(23) = 1.88$; $P = .22$; 4-year-olds: n = 19; 59% correct; $t(18) = 1.99$; $P = .18$; 5-year-olds: n = 14; 54% correct; $t(13) = 0.63$;

$P = .81$, corrected for 3 comparisons). Furthermore, the dominant justifications from succeeders for their correct sorts on the last two test trials were explicit appeals to the words "same" and "different" (62%), when only 10% of non-succeeders justifications were of this type (Fig. 7; see also Appendix B).

The important conclusion from the above analyses is that *some* representations that lead to success on 16-item training trials must differ from those that underlie success on 2-item test trials. At all three ages, there is a decrement in performance from training to test. Most likely, entropy representations are *distinct* from representations of same and different, and *both* are naturally elicited in 16-item AMTS (all-same vs. all-different, and entropy 0 vs. entropy 4), whereas representations of same and different are most likely the sole basis of success on 2-item AMTS. To the extent that the representations underlying 16-item AMTS involve only the contrast between all-same and all-different, then success should transfer without loss to 2-item AMTS, which the above analysis shows is not the case. Conversely, success on 16-item AMTS is clearly not based on entropy alone, at least by age 4, as shown in Experiment 3, in which 4-year-old succeeders matched on the basis of contrasts formulated over representations of same and different (all-same vs. not all-same or all-same vs. all-different).

In sum, Experiments 1–4 together support the following conclusions. First, 3-year-olds rely entirely on entropy in 16-item AMTS in these studies. Experiment 3 showed 3-year-olds to be above chance on this task, and to match intermediate cards to the *all-different* choice card with increasing frequency as entropy increases. They do not succeed at 2-item AMTS, either when trained on 2-item AMTS itself (Experiment S1, and by inference from 4-year-old failure in Experiment 2) or when trained first on 16-item AMTS (Experiment 4). They do not justify their choices by explicit appeal to the words "same" and "different". Thus, representations of same and different seem to be entirely absent from 3-year-olds' approaches to both 16-item and 2-item AMTS. Second, 4-year-olds and 5-year-olds also rely on entropy in 16-item AMTS; their success at this task is even more robust than that of the 3-year-olds is, but for these ages as well, success levels do not generalize freely to 2-item AMTS. However, by age 4, and with increasing frequency by age 5, *some* children spontaneously encode both 16-item all-same and all-different arrays, as well as 2-item AMTS arrays, in terms of the concepts same and different. This encoding is reflected in 100% success on the 2-item AMTS test trials, judgment patterns of intermediate entropy trials that reflect matching based on relations (i.e., those depicted in Fig. 2b and d; Experiment 3), and explicit justifications appealing to the words "same" and "different". Moreover, there were hints that the relations same and different are *more salient* in 16-item arrays than 2-item arrays, as there were numerically more succeeders on 2-item AMTS test trials following training on 16-item arrays, than following training on 2-item arrays, leading to above chance performance in 4-year-olds in Experiment 4, compared to chance performance in Experiment 2.

The above conclusions raise the question of *why* 3-year-olds, and most 4 and 5-year-olds do not use entropy representations to solve 2-item AMTS. Clearly, the distinction between entropy 4 (16-different) and entropy 0 (16-same) is more discriminable than that between entropy 1 (2-different) and entropy 0 (2-same). The animal literature suggests that animals have difficulty discriminating entropy 0 from entropy 1, and the evidence for this is that, having been trained to distinguish 16-same arrays and 16-different arrays, they treat 2-item (and sometimes 4-item) *different* arrays as if they were 16-same arrays (Wasserman & Young, 2010). An alternative (not mutually exclusive) explanation is that entropy simply is not salient or not spontaneously computed in 2-item displays; degree of variability is a more relevant variable in the encoding of larger ensembles than mere pairs. Experiment 5 directly tests whether young children can solve 16-item AMTS when the choice cards and targets differ from each other by entropy values 0 and 1.

6. Experiment 5 – Matching according to entropy 0 vs. 1

In Experiment 5 we ask two questions: first, do 3- and 4-year-old children have the acuity in their representations of entropy to discriminate entropy 0 from entropy 1? If so, they should succeed at a 16-item AMTS task in which the sample and choices have entropy 0 (all-same) or entropy 1 (13-same/3-different). Of course, even if children of a given age can discriminate arrays of these two types, they might do so on the basis of an all-same vs. not all-same rule, rather than on the basis of entropy. As always, we will look at justifications to see whether success is justified by explicit reference to the relations with the words "same" and "different". Second, we ask whether successful training on the 16-same vs. 13-same/3-different transfers to the 2-item AMTS. If 3-year-olds compute only entropy in 16 item arrays, and if their entropy acuity is not up to a 0 vs. 1 comparison, they should fail to distinguish all-same arrays from 13-same/3-different arrays, and of course, continue to fail 2-item AMTS. If the success of some 4-year-olds on 2-item AMTS derives solely from access to the concepts same and different, as suggested above, we would expect that success, if observed, on 13-same/3-different arrays is justified by appeal to the words "same" and "different", and would transfer to 2-item arrays. If, however, the success of 4-year-olds on 2-item AMTS depends on an increasingly acute computation of entropy, we should observe a strong association between success on the 16-same vs. 13-same/3-different task and success on the 2-item AMTS, independently from explicit justifications.

6.1. Methods

6.1.1. Participants

A total of 24 3-year-olds ($M = 3.50$ years; $SD = 0.33$; range = 3.02–3.98 years; 14 girls) and 24 4-year-olds ($M = 4.47$ years; $SD = 0.36$; range = 4.01–4.99 years; 10 girls) participated in this experiment.

6.1.2. Stimuli and procedure

Twenty-four novel all-same 16-item cards, similar to those of Experiment 3, were created. In addition, 24 novel cards were created with 13 same icons and 3 different icons. The procedure is the same as for Experiment 2, except that the experimenter used

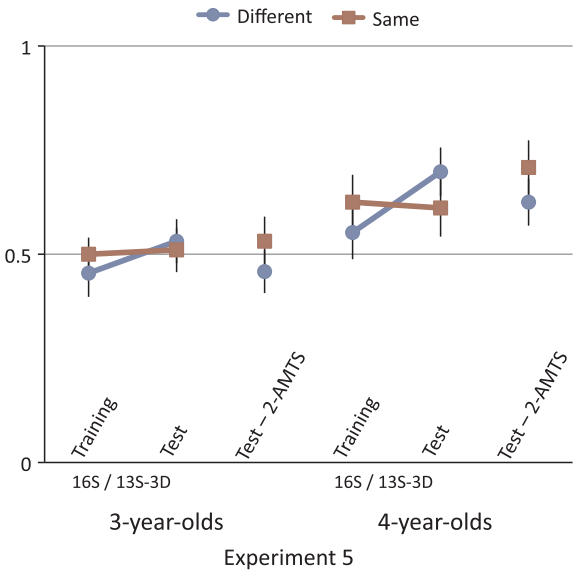

Fig. 10. Results of the training and test in Experiments 5. Error bars represent standard errors from the means.

all-same and 13-same/3-different arrays in training and test. Children first had 8 training trials with feedback and then 8 test trials without feedback, and were asked to explain their choices on the last two test trials. Eight additional test trials with 2-item same- and different-arrays were added at the end of the experiment. The cards and test procedure were identical to those of Experiment 2, except that children were not asked again for justifications.

6.2. Results

The results are analyzed in two steps. First, we examine the performance on the all-same vs. 13-same/3-different AMTS task. We then examine performance on the subsequent 2-item AMTS task. Fig. 10 shows that 4-year-olds could match 16-item arrays of entropy 0 and 1, and generalized success to 2-item AMTS whereas 3-year-olds failed at both. However, as in previous experiments, the group success at age 4 appears to be carried by a small sub-group using explicit representations of same and different. Detailed analyses follow.

6.2.1. 16-item Arrays: training and test

We begin with analyses of performance on the 16-item arrays. A repeated-measure ANOVA with Phase (Training, Test) and Trial Type (all-same, 13-same/3-different) as within-subject factor and Age (3 years, 4 years) as between-subject factor yielded a main effect of Age, as 4-year-olds (62% correct) performed better than 3-year-olds (50% correct); $F(1, 46) = 4.90$; $P = .03$. Other effects and interactions were not significant ($Ps > 0.05$).

Again, we are mainly interested in the age of success. We thus compared performance to chance level at each age. On test trials, 4-year-olds performed better than chance (65% correct, $t(23) = 2.78$; $P = .02$ corrected for 2 comparisons), whereas 3-year-olds were at chance (52% correct, $t(23) = 0.50$; $P = .62$ uncorrected).

In sum, 3-year-olds failed to match 16-item entropy 0 targets (all-same) with the entropy 0 choice and the 16-item entropy 1 targets (13-same/3-different) with the entropy 1 choice. Since they succeeded with entropy 0 vs. entropy 4 arrays (Experiments 3 and 4), this suggests that they are not sensitive to the much smaller entropy contrast probed in this study, which may at least partially explain why they fail 2-item AMTS, which also involves an entropy 0 vs. entropy 1 contrast.

Four-year-olds, in contrast, succeeded. The success of 4-year-olds on the 16-same vs. 13-same/3-different AMTS test trials is carried by 7 succeeders (100%). Non-succeeders, as a group, were at chance on test trials (n = 17; 51% correct; $t(16) = 0.28$; $P = .78$; see Table 1). The failure of the non-succeeders indicates that the 0 vs. 1 entropy contrast is beyond the acuity of most 4-year-olds entropy representations, just as it is for 3-year-olds. Remember that 4-year-olds robustly succeeded at 16-item AMTS in Experiments 3 and 4, when the entropy contrast was 0 vs. 4, and this success was *not* due only to the succeeders.

The 4-year-old *succeeders'* high level of performance on 16-item AMTS in Experiment 5 could be based on the distinction between entropy 0 vs. entropy 1, or it could be based on the distinction between all-same vs. not all-same. An analysis of the justifications supports the latter hypothesis. Fifty-seven percent of the succeeders' justifications used "Same/Different" Language, and a further 14% were One Object justifications (e.g., "these are all bells", "these are all hats"), whereas only 29% were Irrelevant responses (see Fig. 7; Appendix B). Non-succeeders' justifications, in contrast, were mostly (86%) Irrelevant, appealing neither to relations or entropy. The remaining 14% of non-succeeders' explanations for their choices used "Same/Different" language and were produced by 2 children who committed only one error (7/8 correct) on the test trials. This analysis strongly suggests that the performance of 4-year-old succeeders is based on the saliency of the contrast between all-same and not-all same arrays in these 16-item arrays, rather

than on the contrast between entropy 0 and entropy 1.

If the above conclusions are correct, then 3-year-olds should certainly fail the final subsequent 2-item AMTS test, because, by hypothesis, they did not encode the 16-item arrays in terms of the relations same and different and their entropy acuity cannot easily resolve a 1 vs. 0 contrast. Four-year-old succeeders, in contrast, should be above chance on a subsequent 2-item AMTS, because, by hypothesis, they are relying on the distinction between all-same and not all-same and this distinction applies to the 2-item arrays. Four-year-old non-succeeders, like the 3-year-olds, should be at chance, for their entropy acuity, like the 3-year-olds', is not up to a 0 vs. 1 contrast, and they did not encode the arrays in terms of the relations same and different. We now turn to the analyses of the final 2-item AMTS test phase.

6.2.2. 2-item AMTS test trials

A repeated-measure ANOVA with Trial Type (Same, Different) as within-subject factor and Age (3 years, 4 years) as between-subject factor revealed a main effect of Age, 4-year-olds performing better than 3-year-olds; $F(1, 46) = 8.13$; $P = .006$ (see Fig. 10). Other main effects and interactions were not significant (all $Ps > 0.16$). Four-year-olds performed better than chance (67% correct; $t(23) = 3.56$; $P < .01$ corrected for 2 comparisons), whereas, as predicted, 3-year-olds were at chance (49% correct; $t(23) = 0.14$; $P = .89$ uncorrected). As mentioned above, the 3-year-old failure is overdetermined by 3-year-olds' acuity in entropy computations and the total failure, at this age, of matching on the basis of the relations same and different.

Our final analysis contrasted the performance of 4-year-old succeeders and non-succeeders on the 16-same vs. 13-same/3-different AMTS task on the *subsequent* 2-item AMTS task. For this final analysis, the 2 4-year-olds who were correct of 7/8 test trials (one judgment away from succeeder status) were grouped with 8/8 correct participants as succeeders, because, as mentioned above, their justifications patterned with the other succeeders, and not with the non-succeeders.

The performance of the groups of succeeders (n = 7 + 2) and non-succeeders (n = 15) at 16-item AMTS differed on the subsequent 2-item AMTS; $t(22) = 3.30$; $P = .003$. Succeeders were above chance at the subsequent 2-item AMTS task; 83% correct; $t(8) = 4.28$; $P = .003$; non-succeeders were at chance; 57% correct; $t(14) = 1.59$; $P = .13$. Thus, 4-year-olds who could solve the 16-same vs. 13-same/3-different AMTS task generalized their solution to 2-item cards. This behavior is compatible with the rule that they articulate in their explicit justifications on the 16-item AMTS task: all-same vs. not all-same. As in Experiments 3 and 4, most 4-year-olds behaved like the 3-year-olds, failing at the 16-item task, failing to draw on representations of the relations same and different, and unable to rely on entropy, because the entropy contrast of 0 vs. 1 is insufficient to solve the present task.

7. General discussion

The data from Experiments 1–5 support five empirical conclusions and raise several further questions for investigation.

The first conclusion is that the basis of adult success on our16-item AMTS task is representations of the relations same and different rather than entropy, as established by the patterns of matching behavior (those of Fig. 2b and d, rather than that of Fig. 2a), and explicit explanations of the basis of matches.

The second conclusion is that in these AMTS studies, almost all 3-year-olds, and most of 4- and 5-year-olds fail to engage representations of either entropy or the relations same and different on AMTS tasks where the entropy contrast between the choice arrays is 0 vs. 1 (2-item AMTS and 16-same vs. 13-same/3-different), as established by their chance performance and their justifications, which are irrelevant to either entropy or the relations. As Gentner's extensive body of work has established (e.g., Christie & Gentner, 2010; Gentner & Toupin, 1986), this is partly due to a strategy of seeking matches based on individual objects in the arrays.

The third conclusion is that in 16-item AMTS, virtually all 3-year-olds perform better than chance only when they encode the stimuli in terms of entropy. This generalization holds for roughly half of the 4-year-olds as well—the non-succeeders on 16-item AMTS, i.e., those who did not get 8/8 test trials correct and did not appeal to relations in their justifications. These children display all three signatures of animal performance on AMTS tasks that establish that animal success on rule discrimination and AMTS with 16-item arrays is based on graded entropy representations. They were above chance at 16-item AMTS when the entropy difference was 0 vs. 4 (Experiment 3), and failed when it was 0 vs. 1 (Experiment 5). On intermediate arrays, they sorted sample targets with the entropy 4 choice card as a graded function of the entropy of the sample card (Experiment 3). Success on 16-item AMTS did not transfer to 2-item AMTS (Experiment 4). Finally, in all experiments, 3-year-olds were at chance at 2-item AMTS, and most 4-year-olds were as well.

While these signatures make a strong case for entropy, we must acknowledge that we have not run all the controls that have been offered in the animal literature. The animal literature has a much stronger test of entropy representations than those we deploy here, in particular the equivalence of different array mixtures of same and different that lead to the same entropy value, and of course, unequivocal Weber's law assessed thanks to a higher statistical power. Moreover, Wasserman and colleagues have run a number of control experiments to rule out alternative dimensions such as texture and spatial orderliness (reviewed in Wasserman & Young, 2010). The important conclusion we want to reach, however, is that young children, when they succeed in 16-item AMTS, do not use representations of the relations same and different, but rather another representation that is continuous and (at least) correlates with entropy.

The fourth conclusion is that for children, as for adults, the representations that underlie above chance performance at 16-item AMTS are at least partially distinct from those that underlie success on 2-item AMTS (see Castro & Wasserman, 2013 for convergent evidence for the conclusion that 16-item arrays are encoded differently from 2-item arrays in same/different discrimination tasks as well). This is established by the lack of transfer of the level of success from the 16-item task to the 2-item task, and by the fact that 3-year-olds, as well as 4-year-old non-succeeders, achieve better than chance performance on 16-item AMTS in the face of repeated and

abject failure on 2-item AMTS. At the very least, an entropy of 4 is psychologically very different from an entropy of 1, and an entropy of 0 computed from a 2-item array is very different from an entropy of 0 calculated from a 16-item array.

These results are consistent with the possibility that 16-item arrays may elicit spontaneous entropy calculations in these experiments, whereas 2-item arrays may not. In visual working memory, ensemble representations are very different from working memory representations of small sets. Working memory representations of small sets maintain information about each individual in the set (Alvarez & Cavanagh, 2004; Awh, Barton, & Vogel, 2007; Luck & Vogel, 1997; Sperling, 1960; Zhang & Luck, 2008) and have extreme limits on capacity (as a function of the number and complexity of individuals in each set, and the complexity of the working memory model itself—how many sets are being represented in parallel, and their hierarchical organization; e.g., Brady & Alvarez, 2011; Halberda, Sires, & Feigenson 2006). Working memory representations are thus dubbed "parallel individuation models." Ensemble representations, in contrast, do not maintain information about each individual in the set, and are not limited in the number of individuals represented (other than by limits on perceptual resolution of the individuals) but rather capture summary statistics about the individuals all together (e.g., their number, represented by the analog number system – Barth, Kanwisher, & Spelke, 2003; their average size – Ariely, 2001; Chong & Treisman, 2003; Chong & Treisman, 2005a; Chong & Treisman, 2005b; their average orientation – Orban, Vandenbussche, & Vogels, 1984; Parkes, Lund, Angelucci, Solomon, & Morgan, 2001). Furthermore, even if ensemble statistics are computable from small sets, the activation of a parallel individuation working memory model sometimes inhibits the likelihood of creating an ensemble summary statistic (Feigenson, Carey, & Hauser, 2002; Hyde & Spelke, 2009; Hyde & Wood, 2011).

Entropy is a property of an ensemble of individuals (being a measure of the degree of variation among them), whereas same and different are relations between individuals. Sixteen-item arrays are canonical ensembles; 2-item arrays are canonical sets that elicit parallel individual working memory models. This fact alone may explain why the two types of arrays elicit distinct representations. Furthermore, even if the 2-item arrays do elicit the computation of ensemble statistics, the entropy representations of young children do not have the resolution to distinguish entropy 0 from entropy 1 (Experiment 5).

The fifth conclusion is that the developmental emergence of robust success on 2-item AMTS between ages 3 and 6, asserted by Premack (1983) and confirmed here, is due to increased availability of explicit representations of same and different as a basis for matching sample to choice cards over these years. In Experiment 2, we found that the success of 5-year-olds was due to a small group (29%) who immediately and spontaneously solved the 2-item AMTS task, performing statistically as well on the training trials as on the test trials, and performing perfectly on the test trials (succeeders), whereas the rest (71%) were at chance on both training and test. Indeed, in every age group on every 2-item AMTS, children either performed perfectly during test, even succeeding statistically on the training trials, or they were at chance throughout training and test. An analysis of children's justifications showed that those who succeeded, like adults in Experiment 1, matched on the basis of the distinction between the concepts *same* and *not same* or *different*, as expressed by the words "same" and "different".

The analyses of justifications presented in the text, and expanded in Appendix B, established that succeeders in any entropy 0 vs. entropy 1 contrast were far more likely to justify their choices by appeals to the relations same and different than were non-succeeders, who virtually never did so. In Appendix C, we present the converse analysis—dividing children according to whether they gave at least one "Same/Different" Language or All One Object justification, or only irrelevant justifications. At all ages, in all entropy 0 vs. entropy 1 contrasts, children who offered one of the former two types of justifications for at least one of their two probed justifications, performed as a group well over 85% correct on the 8 test trials, whereas those who provided only irrelevant justifications performed at chance (see Table C1). Thus, the association between explicit verbal access to the relations same and different and success on 2-item AMTS, as well as to 16-item all-same vs. 13-same/3-different AMTS, is very strong. Part of what is changing developmentally, between ages 3 and 6, is the spontaneous availability of explicit representations of concepts *same* and *different* as bases for establishing matches in both 16-item and 2-item AMTS.

These results converge with Christie and Gentner's (2014) findings that children's success at matching AA to XX is facilitated by a previous task eliciting the use of the words "same" and "different", but only for those children who appear to understand these words (3- and 4-year-olds). There are two possibilities as to what the words "same" and "different" mean for young children. One possibility is that children learn to categorize the continuous dimension of entropy into two categories: 0 and more than 0. According to this view, "same" means entropy 0 and "different" means entropy higher than 0 (as suggested by Wasserman and Young (2010) even for adults). To explain the results of our 5 experiments in terms of this possibility, we must also assume that the acuity of the entropy representation increases with age, so that between 3 and 6, children become able to discriminate entropies of 0 and 1. Of course, becoming able to discriminate entropies of 0 and 1 is a necessary prerequisite for applying a categorical distinction between 0 and not 0 to the entropies 0 and 1, but it is not a sufficient condition for establishing that categorical distinction. Improvement of acuity and drawing a categorical distinction would have to work hand in hand to yield the ability to solve the 2-item AMTS task, generalize from 16-item AMTS to 2-item AMTS and categorize any arrays of non-zero entropy together.

The second possibility is that the words "same" and "different" do not refer to representations of entropy, but rather to the relations same and different. As discussed above, one key distinction between entropy and relations is that entropy is a property of an ensemble of items, whereas relations are relations between items (or between properties). There is a large and complex literature investigating the semantics of the words "same" and "different" (e.g., Barker, 2007; Beck 2000; Carlson 1987; Dowty 1985; Heim 1985; Moltmann 1992). While all debates are far from being resolved, all authors agree that, for adults, "same" and "different" refer to relations (2-place predicates), not to entropy (a 1-place predicate). Indeed, the words "same" and "different" behave in many situations like comparatives (which are by definition relations between two things that are being compared; Alrenga, 2007; Charnavel, 2015; Oxford, 2010). Most importantly, like comparatives, "same" and "different" can occur with comparative clauses (1), which constitute the hallmark grammatical property of a comparative construction (Oxford, 2010).

(1) a. Sue gave [as good an/the same] answer as I expected.
b. Sue gave a [better/different] answer than I expected.

Other evidence that for adults the meanings of the words "same" and "different" express relations rather than the distinction between entropy 0 and entropy higher than 0 is the way the meanings compose with quantifiers, in phrases like "all the same as each other", or "all different from each other." These analyses, of course, are concerned with the semantics of "same" and "different" for adults. Do these words assume the same semantics for young children?

Children come to understand that the words "same" and "different" refer respectively to pairs of same and different elements in the fourth year of life (Webb et al., 1979). Data from our lab confirms that only half of 3-year-olds comprehend the words "same" and "different," whereas virtually all English learners age 4 or older do so (Hochmann et al., unpublished data). Interestingly, the results of Experiment 5 suggest that the acuity of the representation of entropy is too low at these ages to discriminate between 0 and 1. Thus, 50% of 3-year-olds understand the words "same" and "different", whereas none could match sets of entropies 0 and 1; and all 4-year-olds understand the words "same" and "different", whereas only at most 37% of them could match sets of entropies 0 and 1 (Experiment 5). While this argument is insufficient to conclude that "same" and "different" refer to relations in preschool children, it certainly suggests that "same" and "different" do not refer to the contrast between 0 and non-0 entropy when they are first learned, for these words can be correctly assigned to arrays whose entropies the child cannot resolve. Further studies exploring the semantics of the meanings of "same" and "different" for 3- and 4-year-old children should explore how the meanings of these words combine with those of quantifiers and logical connectives (Barker, 2007). Ongoing work in our lab is exploring whether children this young can compose meanings, for the expressions "all the same" and "not all the same" for example, as soon as they learn the words "same" and "different." If they can, and can apply these meanings to the 16-item all-same arrays vs. the 13 same-3 different arrays, before they can distinguish entropy 0 and 1, it would provide strong evidence that these words express relations, not entropy, as soon as they are learned.

These results raise the possibility that the representations underlying the meanings of the linguistic symbols "same" and "different" may differ in some ways from the representations of same and different that underlie non-human animal success on 2-item AMTS (Fagot & Parron, 2010; Fagot & Thompson, 2011; Obozova et al., 2015; Smirnova et al., 2015), same/different discrimination (e.g., Thompson et al., 1997; Wasserman & Young, 2010) and simple match to sample (e.g., Giurfa et al., 2001), as well as differing in some ways from the representations of same and different that underlie infant success on match to sample (Hochmann et al., 2016), same/different discrimination (Hochmann, 2010; Kovács, 2014; Walker & Gopnik, 2014), and habituation to same and different (Addyman & Mareschal, 2010; Ferry et al., 2015). As discussed above, some of these results could be explained by representations of symmetry. Moreover, Hochmann et al. (2016) suggested that the pre-lexical representation of same may not consist in a unitary symbol, but rather in a variable repetition. Furthermore, in several rule learning (Hochmann, 2010; Hochmann et al., 2011; Kovács, 2014) and MTS studies (Hochmann et al., 2016; see also Zentall et al., 1981, for a related finding in pigeons), infants succeeded on the basis of the representation of same alone, suggesting that they cannot easily compose *same* with *not* to create a representation with the content *different*. Thus, one change with development might be the creation of an explicit summary symbol "same," which composes easily with a representation of "not" to yield "not same" = "different." Having single summary labels ("same" and "different," where "different" = "not same") could well help on the AMTS task, as these symbols can be stored in a working memory representation of the sample card and applied to the choice cards, turning the task into a simple match to sample task.

However, *comprehending* the words is clearly not sufficient for adult-like performance, as seen by the small percentage of 4- and 5-year-olds that respond like adults on 2-item AMTS in the present experiments. That is, children who know the words "same" and "different" and can reliably apply them to pairs of stimuli that are the same vs. pairs of stimuli that are different presumably have the capacity to represent the sample cards and the choice cards in terms of the relations that establish the relevant matches. Something changes between ages 4 and 6 that makes it increasingly more *likely* that they do so spontaneously in these studies. Even if learning the linguistic summary symbols "same" and "different" involves some change in the underlying representations of the corresponding relations, such a representational change is not the whole story of what is changing between ages 3 and 6.

What else might be changing over these years, in addition to mastering the language of referring to relations? Gentner and her colleagues have provided evidence that younger children must overcome a bias to entertain hypotheses concerning similarity among the individual objects in the array (the object bias), and also provided evidence that progressive alignment involving several examples of pairs of entities that instantiate the relation can induce an increase of saliency of the relational hypothesis, both in infants (Ferry et al., 2015) and in young children (Christie & Gentner, 2014). Some developmental process must make the hypothesis that the relations same and different are relevant to the task at hand *more salient*, such that it is almost the first hypothesis that comes to mind in these AMTS tasks (remember succeeders succeeded with virtually no training, other than the admonition if they erred on the first trials "to look at the whole card").

Several factors may participate in making the relational hypothesis more salient, and future studies should try to investigate their respective roles. First, it has long been observed that perception is hierarchical (e.g., a forest is composed of trees, which possess leaves, and so on). The relative importance of local (e.g., focusing on the trees) and global (e.g., focusing on the forest) visual processing appears to shift between 3 and 6 years of age (Dukette & Stiles, 1996; Nayar, Franchak, Adolph, & Kiorpes, 2015; Poirel, Mellet, Houdé, & Pineau, 2008), as a consequence of brain maturation (Fink et al. 1996; Poirel et al., 2011). While global processing is usually described as establishing spatial relations to form a global structure, it is not impossible that similar processes are at play in detecting abstract relations such as same and different. Interestingly, young children's difficulty with global processing is greater when the set contains few elements than when it contains many elements (Kimchi, Hadad, Behrmann, & Palmer, 2005), possibly explaining why the processing of a global property such as entropy for large ensembles in young children is not problematic, but processing relational properties of 2-item arrays is.

Second, the development of executive functions (Anderson, 2002; Diamond, 2013; Garon, Bryson, & Smith, 2008; Kharitonova & Munakata, 2011; Welsh, Pennington, & Groisser, 1991; Zelazo, Carter, Reznick, & Frye, 1997), particularly inhibitory skills, may help children coordinate representations of the individual and of the array, and inhibit the object bias. Young children may well consider the relational hypothesis in the 2-item AMTS, but they might have trouble resisting the appeal of the object-level matching hypothesis. Evidence that progressive alignment (Christie & Gentner, 2014) helps children is consistent with this hypothesis.

Finally, the capacity for abstraction also develops between 3 and 6 years of age, a fact that has also been related to executive functions development (Kharitonova & Munakata, 2011), perhaps rendering more salient abstract relational concepts such as *same* and *different*. The above three suggestions are neither mutually exclusive nor necessarily independent from each other.

8. Conclusion

The 2-item AMTS, classically called relational match-to-sample task, is notoriously extremely difficult for non-human species. Three- and most 4-year-old children, likewise, struggle with the task. Like baboons and pigeons, they fail to match pairs of pictures on the basis of their same/different relations. Like baboons and pigeons, they however succeed when arrays of 16 pictures are used instead of pairs. Like baboons and pigeons, their behavior can be explained by a property of ensembles: entropy. In contrast, a few 4-year-olds and about half of 5-year-olds spontaneously solve the classic relational match-to-sample. This development appears due to two factors: (1) the recruitment of representations corresponding to the meaning of the words “same” and “different” (most likely relations), and (2) the increasing saliency of these representations, possibly linked with the development of global processing, executive functions and abstraction.

Acknowledgments

This research was supported by the project grant ANR-16-CE28-0006-01 awarded by the Agence Nationale pour la Recherche – France to J-R.H. and the NIH – United States Research Project Grant R01-HD038338 awarded to S.C.

Appendix A. Experiment S1 – 2-item array match to sample

Experiment S1 presents children with a problem they could solve by learning a same/different discriminative rule “Doggy likes same-cards; Panda likes different-cards” or by learning full 2-item AMTS rule: matching X X to A A and X Y to C D.

A.1. Methods

A.1.1. Participants

One hundred children were tested in this experiment, of which 92 were recruited as they were visiting Boston Children’s museum with their parents and 8 (five 3-year-olds and three 5-year-olds) were tested in the laboratory. These families were middle class, and over 80% non-Hispanic Caucasian, with the rest being Hispanic, black or Asian. Children were divided in four age groups: 3-year-olds (3y 0 m–3y 11 m; N = 26), 4-year-olds (4y 0 m–4y 11 m; N = 31), 5-year-olds (5y 0 m–5y 11 m; N = 28) and 6-year-olds (6y 0 m–6y 11 m; N = 15).

A.1.2. Stimuli

We selected 92 different symbols to create 81 cards, each displaying 2 symbols. Forty were same-cards, 13 of which followed a vertical arrangement to serve as samples, and 27 followed a diagonal arrangement to serve as choice cards. Forty-one were different-cards, 13 of which followed a vertical arrangement to serve as samples, and 28 a diagonal arrangement to serve as choice cards. Twenty-four sample cards (12 same-cards and 12 different-cards), 24 same-cards and 24 different-cards were randomly selected for each participant.

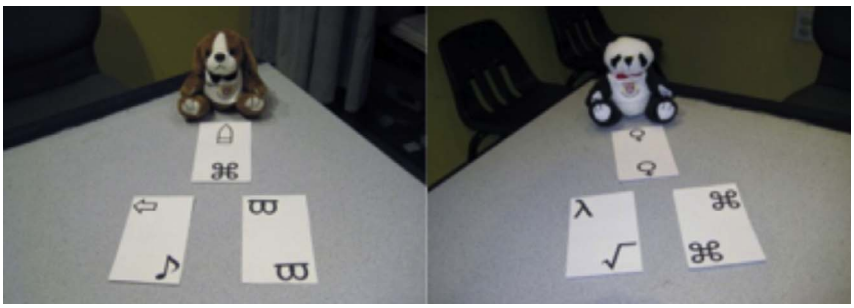

Fig. A1. Example of what children saw in one same-trial (left) and in one different-trial (right).

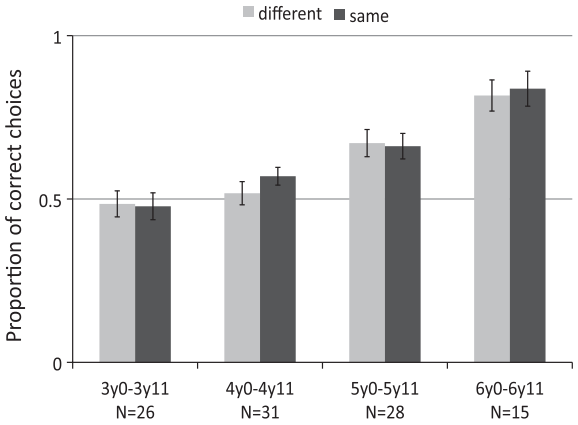

Fig. A2. Results of Experiment 1. Dark grey bars represent the proportions of correct responses when the correct choice was the puppet who liked the same-cards; and light grey bars represent the proportions of correct responses when the correct choice was the puppet who liked the different-cards. Error bars represent standard errors from the mean.

A.1.3. Procedure

At the beginning of the experiment, an experimenter announced to the child that they were going to play a game. She introduced two puppets, Doggy and Panda, and placed them on the table on the left and right of the experimenter. "Each puppet," she said, "likes a certain type of cards, and giving them a card they like makes them happy. When they are happy, they give the child a sticker as reward." Then, on each trial (Fig. A1), the experimenter showed a card (the sample card) and placed it in front of one of the puppet (e.g., Doggy). The experimenter attracted the child's attention to the card and announced that Doggy (or Panda) likes it. She then placed two other cards (the choice cards) in front of the sample and asked: "if Doggy/Panda likes this card [pointing at the sample], which of these cards [pointing at the two choice cards] would he also like?"

The sample was a same-card in half of the trials, and a different-card in the other half. Either Panda liked the same-cards and DoggyDoggie the different-cards, or vice versa, counter-balanced across children. The order of Panda trials and DoggyDoggie trials was randomized, subject to the constraint that no more than three trials of the same type (e.g., Panda-same) could be presented in a row. The task could be solved either by doing a relational match-to sample, or by learning two rules such as "Panda like same-cards and Doggy likes different-cards".

If the child chose correctly by touching, pointing at or handing the correct card, he was given a sticker as a reward. The sticker was placed in a small bag, for the child to retrieve at the end of the study. If the child chose incorrectly, the experimenter showed what the correct choice was and placed the correct card next to sample, attracting the child's attention to the two cards. Children's responses were coded offline. Due to the difficulties in testing in a Children's museum, there was some variability in the number of trials. Seventy-six children completed 24 trials, 15 completed 20 trials, and 9 children completed only 12–19 trials.

A.2. Results

The results, shown in Fig. A2, confirm the results reported by Premack (1983): children begin to succeed in 2-item AMTS at 5 years of age, but not earlier. Detailed analyses follow.

The proportion of correct responses was computed considering all trials but separating same-trials and different-trials (Fig. A2). A repeated-measure ANOVA with Trial Type (Same, Different) as within-subject factor and Age (3 years, 4 years, 5 years, 6 years) as between-subject factor revealed a main effect of Age; $F(3, 96) = 15.42$; $P < .001$. There were no significant main effects or interactions involving trial type. Bonferroni post hoc tests showed that 3- and 4-year-olds did not differ from each other ($P = .17$), but each differed from both 5- and 6-year-olds ($Ps < .007$). Five- and 6-year-olds further differed from each other ($P = .025$).

The above ANOVA establishes that children improve equivalently with same-trials and different-trials over the ages of 3–6 on this task. Our interest, of course, is when they actually succeed on the task. Therefore, we compared children's performance at each age to chance, collapsing over same- and different-trials. In this analysis, and in every analysis comparing performance levels to chance in this paper, we carry out a conservative test for multiple comparisons (to minimize the likelihood that 5-year-olds, for example, perform better than chance, by chance). That is, if there are N age groups, we use a Bonferroni adjustment, correcting for N comparisons (unless the corrected p-value exceeds 1). Three-year-olds and 4-year-olds were at chance (respectively 48% correct, $t(25) = 0.59$; $P = .56$ uncorrected; and 54% correct, $t(30) = 1.67$; $P = .44$ corrected for 4 comparisons). Five-year-olds performed better than chance (67% correct; $t(27) = 4.46$; $P < .001$), as did 6-year-olds (83% correct; $t(14) = 6.67$; $P < .0001$, both corrected for 4 comparisons).

In sum, Experiment 1 confirms Premack's claim that children can solve the 2-item AMTS by age 5, but not robustly until age 6.

Appendix B

The analyses in the text reporting the justifications as a function of succeeder-levels (8 out of 8 correct vs. fewer than 8 out of 8 correct on test trials) collapsed over age, because at all ages non-succeeders as a group performed at chance on test trials, and because, as the analyses showed, the justifications showed the two groups to be qualitatively different from each other (see also Appendix C). Fig. B1 displays the distribution of justifications as a function of succeeder level for *each* sample of 24 participants in every condition.

What one sees from Fig. B1 is that at all ages, in all entropy 0–1 conditions (2-item AMTS: Experiments 2, 4; 16-item AMTS: Experiment 5), non-succeeders overwhelming gave irrelevant justifications. Of course, succeeders, especially the younger ones, also sometimes gave irrelevant justifications, but the 50% irrelevant justifications by 4-year-old succeeders in Experiment 2 is misleading as there only was one succeeder at this age, who said "because this is that [point to top picture of one card] and I don't know what this is [point to the other picture on the card]" for one of her 2 justifications; similarly, there was only 1 3-year-old succeeder in Experiment 5. And as can be seen from this graph, the relevant justifications, provided virtually only by succeeders at every age and in every 0–1 entropy contrast condition, were almost always explicit references to the relations same and different, and secondarily references to the fact that the sample was all x's (e.g., all hammers) and the choice was all y's (e.g., all bells), along sometimes with a comment that the sample or the other choices was not all z's.

Fig. B1 further confirms that the very different pattern observed in Experiment 3 (16- item AMTS; entropy contrast 0 vs. 4) that was reported in the text holds at both ages 3 and 4. Now many succeeders provide irrelevant justifications, at both ages tested.

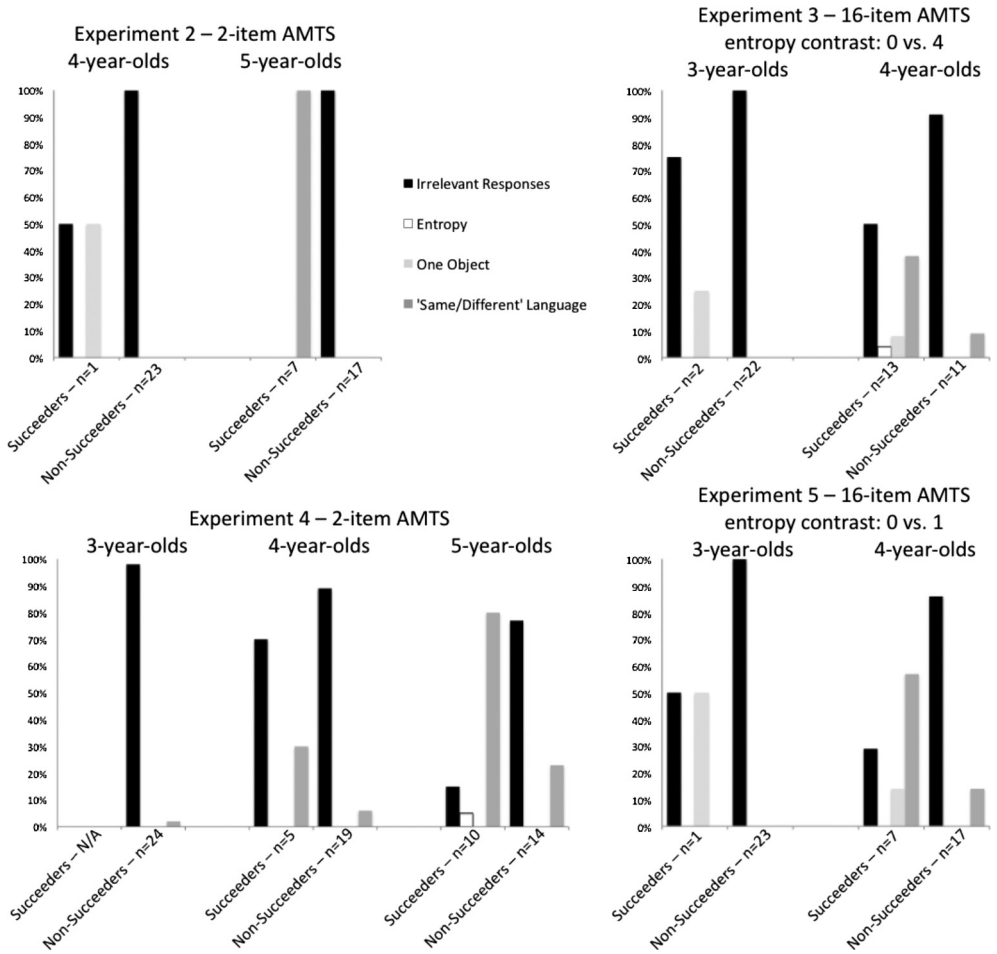

Fig. B1. Distribution of the different types of justifications for the last two AMTS tests in Experiments 2–5. In all experiments, succeeders are those participants who made no error in the test ($P < .01$; 2-tail binomial test), whereas non-succeeders made at least one error ($P > .07$; 2-tail binomial test).

Appendix C

The analyses of justifications in the text showed that succeeders were more likely than non-succeeders to justify their choices by appeal to relations ("Same/Different" Language justifications—appeals to the relations same and different; One Object Justifications–appeals to the contrast between all xs and not all ys; see also Appendix B). The analysis summarized in Table C1 shows that the converse is also true. Only those participants who provided relational justifications were, as a group, above chance at 2-item AMTS (Experiments 2, 4, and 5; in Experiment 5 the justifications were for test trials on the prior 16-same/13s-3d 16-item AMTS test trials). Also, only those children who provided Relational justifications on the 16-same/13s-3d 16-item AMTS test trials were above chance on those very test trials—(Table C1). That is, only children who explicitly appeal to the relations same and different when explaining their choices performed above chance on AMTS where the entropy contrast was 0/1. This was true at every age, except for age 3 in Experiments 4 and 5, where only 1 child per experiment explicitly appealed to the relations same and different, although these children also were over 62% correct overall on the test trials in the entropy 0 vs. entropy 1 conditions—Experiments 4 and 5). In Experiment 3, in contrast, on 16-item all-same/all-different AMTS (entropy 0 vs. entropy 4), even children who did not appeal to the relations same and different, including 3-year-olds, were above chance as a group. This analysis provides further support for the conclusion that children can solve the 16-item AMTS using the entropy contrast between 0 and 4 in Experiment 3, whereas they need to rely on relational representations in all other versions of AMTS used in this study (entropy contrasts 0 vs 1).

Table C1
Children's behavior on test trials according to their justifications.

	At least one *'Same/Different' Language* or *One Object* justification			Only *Irrelevant* or *Entropy* justifications		
	Age			Age		
	3	4	5	3	4	5
Exp 2 2-item	–	–	100%** (N = 7)	–	54%ns (N = 24)	60%ns (N = 17)
Exp 3 16-item entropy 0/4	100% (N = 1)	97%** (N = 9)	–	65%** (N = 23)	80%** (N = 15)	–
Exp 4 2-item	62% (N = 1)	87%* (N = 4)	90%** (N = 12)	54%ns (N = 23)	62%ns (N = 20)	56%ns (N = 12)
Exp 5 16-item entropy 0/1	100% (N = 1)	97%** (N = 8)	–	50%ns (N = 23)	50%ns (N = 16)	–
Exp 5 2-item	100% (N = 1)	89%** (N = 8)	–	47%ns (N = 23)	55%ns (N = 16)	–

** $P < .01$.
* $P < .05$.
ns $P > .05$.

Appendix D. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cogpsych.2017.11.001.

References

Addyman, C., & Mareschal, D. (2010). The perceptual origins of the abstract same/different concept in human infants. *Animal Cognition, 13*(6), 817–833.
Alrenga, P. (2007). *Dimensions in the semantics of comparatives.* PhD thesisUniversity of California Santa Cruz.
Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science, 15*(2), 106–111.
Anderson, P. (2002). Assessment and development of executive function (EF) during childhood. *Child Neuropsychology, 8*(2), 71–82.
Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science, 12*(2), 157–162.
Awh, E., Barton, B., & Vogel, E. K. (2007). Visual working memory represents a fixed number of items regardless of complexity. *Psychological Science, 18*(7), 622–628.
Barker, C. (2007). Parasitic scope. *Linguistics and Philosophy, 30*, 407–444.
Barth, H., Kanwisher, N., & Spelke, E. (2003). The construction of large number representations in adults. *Cognition, 86*, 201–221.
Beck, S. (2000). The semantics of different: Comparison operator and relational adjective. *Linguistics and Philosophy, 23*(2), 101–139.
Blough, D. S. (1959). Delayed matching in the pigeon. *Journal of the Experimental Analysis of Behavior, 2*(2), 151–160.
Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science, 22*(3), 384–392.
Carey, S. (2009). *The origin of concepts.* New York: Oxford University Press.
Carlson, G. (1987). Same and different: Some consequences for syntax and semantics. *Linguistics and Philosophy, 10*, 531–565.

Castro, L., & Wasserman, E. A. (2013). Humans deploy diverse strategies in learning same-different discrimination tasks. *Behavioural Processes, 93*, 125–139.
Castro, L., Young, M. E., & Wasserman, E. A. (2006). Effects of number of items and visual display variability on same-different discrimination behavior. *Memory & Cognition, 34*(8), 1689–1703.
Charnavel, I. (2015). Same, different and other as comparative adjectives - A uniform analysis based on French. *Lingua, 156*, 129–174.
Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research, 43*, 393–404.
Chong, S. C., & Treisman, A. (2005a). Attentional spread in the statistical processing of visual displays. *Perception and Psychophysics, 67*(1), 1–13.
Chong, S. C., & Treisman, A. (2005b). Statistical processing: Computing the average size in perceptual groups. *Vision Research, 45*(7), 891–900.
Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development, 11*, 356–373.
Christie, S., & Gentner, D. (2014). Language helps children succeed on a classic analogy task. *Cognitive Science, 38*(2), 383–397.
Cook, R. G., & Wasserman, E. A. (2007). Learning and transfer of relational matching-to-sample in pigeons. *Psychonomic Bulletin and Review, 14*(6), 1107–1114.
Delius, J. D., & Nowak, B. (1982). Visual symmetry recognition by pigeons. *Psychological Research Psychologische Forschung, 44*(3), 199–212.
Diamond, A. (2013). Executive functions. *Annual Review of Psychology, 64*, 135–168.
Dowty, D. (1985). *A unified indexical analysis of same and different: A Response to Stump and Carlson.* Unpublished manuscriptThe Ohio State University.
Dukette, D., & Stiles, J. (1996). Children's analysis of hierarchical patterns: Evidence from a similarity judgment task. *Journal of Experimental Child Psychology, 63*(1), 103–140.
Fagot, J., & Parron, C. (2010). Relational matching in baboons (Papio papio) with reduced grouping requirements. *Journal of Experimental Psychology: Animal Behavior Processes, 36*(2), 184.
Fagot, J., & Thompson, R. K. (2011). Generalized relational matching by guinea baboons (Papio papio) in two-by-two-item analogy problems. *Psychological Science, 22*(10), 1304–1309.
Fagot, J., Wasserman, E. A., & Young, M. E. (2001). Discriminating the relation between relations: the role of entropy in abstract conceptualization by baboons (Papio papio) and humans (Homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes, 27*(4), 316.
Fechner, G. T. (1966/1860). *Elements of psychophysics* (H. E. Adler, Trans.) (Vol. 1). New York: Rinehart & Winston.
Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object-files versus analog magnitudes. *Psychological Science, 13*, 150–156.
Ferry, A. L., Hespos, S. J., & Gentner, D. (2015). Prelinguistic relational concepts: Investigating analogical processing in infants. *Child development, 86*(5), 1386–1405.
Fink, G. R., Halligan, P. W., Marshall, J. C., Frith, C. D., Frackowiak, R. S. J., & Dolan, R. J. (1996). Where in the brain does visual attention select the forest and the trees? *Nature, 382*(6592), 626–628.
Garon, N., Bryson, S. E., & Smith, I. M. (2008). Executive function in preschoolers: A review using an integrative framework. *Psychological Bulletin, 134*(1), 31.
Gentner, D. (2003). Why we're so smart. In D. Gentner, & S. Goldin-Meadow (Eds.). *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.
Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*, 277–300.
Giurfa, M., Eichmann, B., & Menzel, R. (1996). Symmetry perception in an insect. *Nature, 382*, 458–461.
Giurfa, M., Zhang, S., Jenett, A., Menzel, R., & Srinivasan, M. V. (2001). The concepts of 'sameness' and 'difference' in an insect. *Nature, 410*(6831), 930–933.
Halberda, J., Sires, S. F., & Feigenson, L. (2006). Multiple spatially-overlapping sets can be enumerated in parallel. *Psychological Science, 17*(7), 572–576.
Harley, H. E., Putman, E. A., & Roitblat, H. L. (2003). Bottlenose dolphins perceive object features through echolocation. *Nature, 424*(6949), 667–669.
Heim, I. (1985). *Notes on comparatives and related matters*. Ms, University of Texas, Austin < www.semanticarchive.net > .
Hochmann, J.-R. (2010). *Categories, words and rules in language acquisition.* PhD dissertationTrieste: SISSA.
Hochmann, J.-R., Benavides-Varela, S., Fló, A., Nespor, M., & Mehler, J. (2017). Bias for vocalic over consonantal information in 6-month-olds. *Infancy*. http://dx.doi.org/10.1111/infa.12203.
Hochmann, J.-R., Benavides-Varela, S., Nespor, M., & Mehler, J. (2011). Vowels and consonants in early language acquisition. *Developmental Science, 14*, 1445–1458.
Hochmann, J-R., Carey, S., & Mehler, J. (submitted for publication). Same but not different? Representations of abstract relations in infancy.
Hochmann, J.R., Zhu, R., & Carey, S. (unpublished). Acquisition of the words 'same' and 'different' in preschoolers' vocabulary.
Hochmann, J.-R., Mody, S., & Carey, S. (2016). Infants' representations of same and different in match- and non-match-to-sample tasks. *Cognitive Psychology, 86*, 87–811.
Hyde, D. C., & Spelke, E. S. (2009). All numbers are not equal: An electrophysiological investigation of small and large number representations. *Journal of Cognitive Neuroscience, 21*(6), 1039–1053.
Hyde, D. C., & Wood, J. N. (2011). Spatial attention determines the nature of non-verbal numerical cognition. *Journal of Cognitive Neuroscience, 23*(9), 2336–2351.
James, W. (1890/1950). The principles of psychology (Vol. 1). Dover Publications.
Kharitonova, M., & Munakata, Y. (2011). The role of representations in executive function: Investigating a developmental link between flexibility and abstraction. *Frontiers in Psychology, 2*, 1–12.
Kimchi, R., Hadad, B., Behrmann, M., & Palmer, S. E. (2005). Microgenesis and ontogenesis of perceptual organization evidence from global and local processing of hierarchical patterns. *Psychological Science, 16*(4), 282–292.
Kovács, Á. M. (2014). Extracting regularities from noise: Do infants encode patterns based on same and different relations? *Language Learning, 64*(s2), 65–85.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279–281.
Moltmann, F. (1992). Reciprocals and same/different: Towards a semantic analysis. *Linguistics and Philosophy, 15*(4), 411–462.
Mumby, D. G. (2001). Perspectives on object-recognition memory following hippocampal damage: Lessons from studies in rats. *Behavioural Brain Research, 127*(1), 159–181.
Nayar, K., Franchak, J., Adolph, K., & Kiorpes, L. (2015). From local to global processing: The development of illusory contour perception. *Journal of Experimental Child Psychology, 131*, 38–55.
Obozova, T., Smirnova, A., Zorina, Z., & Wasserman, E. (2015). Analogical reasoning in amazons. *Animal Cognition, 18*(6), 1363–1371.
Oden, D. L., Thompson, R. K., & Premack, D. (1988). Spontaneous transfer of matching by infant chimpanzees (Pan troglodytes). *Journal of Experimental Psychology: Animal Behavior Processes, 14*(2), 140.
Orban, G. A., Vandenbussche, E., & Vogels, R. (1984). Meridional variations and other properties suggesting that acuity and orientation discrimination rely on different neuronal mechanisms. *Ophthalmic and Physiological Optics, 4*(1), 89–93.
Oxford, W. (2010). *Same, other, and different: A first look at the microsyntax of identity adjectives.* [PhD thesis]University of Toronto.
Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience, 4*(7), 739–744.
Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*, 109–178.
Poirel, N., Mellet, E., Houdé, O., & Pineau, A. (2008). First came the trees, then the forest: Developmental changes during childhood in the processing of visual local-global patterns according to the meaningfulness of the stimuli. *Developmental Psychology, 44*(1), 245.
Poirel, N., Simon, G., Cassotti, M., Leroux, G., Perchey, G., Lanoë, C., ... Houdé, O. (2011). The shift from local to global visual processing in 6-year-old children is associated with grey matter loss. *PLoS ONE, 6*(6), e20879.
Premack, D. (1983). The codes of man and beast. *Behavioral and Brain Sciences, 6*, 125–137.
Schluessel, V., Beil, O., Weber, T., & Bleckmann, H. (2014). Symmetry perception in bamboo sharks (Chiloscyllium griseum) and Malawi cichlids (Pseudotropheus sp.). *Animal Cognition, 17*(5), 1187–1205.
Smirnova, A., Zorina, Z., Obozova, T., & Wasserman, E. (2015). Crows spontaneously exhibit analogical reasoning. *Current Biology, 25*(2), 256–260.
Spelke, E. S. (2003). What makes us smart? Core knowledge and natural language. In D. Gentner, & S. Goldin-Meadow (Eds.). *Language in mind: Advances in the study of language and thought*. Cambridge, MA: MIT Press.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1–29.
Thompson, R. K. R., & Oden, D. L. (1995). A profound disparity revisited: Perception and judgement of abstract identity relations by chimpanzees, human infants and monkeys. *Behavioral Processes, 35*, 149–161.
Thompson, R. K., Oden, D. L., & Boysen, S. T. (1997). Language-naïve chimpanzees (Pan troglodytes) judge relations between relations in a conceptual matching-to-sample task. *Journal of Experimental Psychology: Animal Behavior Processes, 23*(1), 31–43.
Tyrrell, D. J., Stauffer, L. B., & Snowman, L. G. (1991). Perception of abstract identity/difference relationships by infants. *Infant Behavior and Development, 14*(1), 125–129.
von Fersen, L., Manos, C. S., Goldowsky, B., & Roitblat, H. (1992). Dolphin detection and conceptualization of symmetry. *Marine mammal sensory systems* (pp. 753–762). US: Springer.
Vonk, J. (2003). Gorilla (Gorilla gorilla gorilla) and orangutan (Pongo abelii) understanding of first-and second-order relations. *Animal Cognition, 6*(2), 77–86.
Walker, C. M., Bridgers, S., & Gopnik, A. (2016). The early emergence and puzzling decline of relational reasoning: Effects of knowledge search on inferring abstract concepts. *Cognition, 156*, 30–40.
Walker, C. M., & Gopnik, A. (2014). Toddlers infer higher-order relational principles in causal learning. *Psychological Science, 24*, 87–92.
Wasserman, E. A., Fagot, J., & Young, M. E. (2001). Same-different conceptualization by baboons (Papio papio): the role of entropy. *Journal of Comparative Psychology, 115*, 42–52.
Wasserman, E. A., & Young, M. E. (2010). Same–different discrimination: The keel and backbone of thought and reasoning. *Journal of Experimental Psychology: Animal Behavior Processes, 36*(1), 3.
Wasserman, E. A., Young, M. E., & Fagot, J. (2001). Effects of number of items on the baboon's discrimination of same from different visual displays. *Animal Cognition, 4*, 163–170.
Wasserman, E. A., Young, M. E., & Nolan, B. C. (2000). Display variability and spatial organization as contributors to the pigeon's discrimination of complex visual stimuli. *Journal of Experimental Psychology: Animal Behavior Processes, 26*, 133–143.
Waxman, S. R., & Braun, I. (2005). Consistent (but not variable) names as invitations to form object categories: new evidence from 12-month-old infants. *Cognition, 95*(3), B59–B68.
Webb, R. A., Oliveri, M. E., & O'Keeffe (1979). Investigations of the meaning of "different" in the language of young children. *Child Development, 45*(4), 984–991.
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*(2), 131–149.
Wright, A. A., Cook, R. G., Rivera, J. J., Sands, S. F., & Delius, J. D. (1988). Concept learning by pigeons: Matching-to-sample with trial-unique video picture stimuli. *Animal Learning & Behavior, 16*(4), 436–444.
Young, M. E., & Wasserman, E. A. (1997). Entropy detection by pigeons: Response to mixed visual displays after same-different discrimination training. *Journal of Experimental Psychology: Animal Behavior Processes, 23*, 157–170.
Young, M. E., & Wasserman, E. A. (2001). Entropy and variability discrimination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(1), 278.
Young, M. E., Wasserman, E. A., & Garner, K. L. (1997). Effects of number of items on the pigeon's discrimination of same from different visual displays. *Journal of Experimental Psychology: Animal Behavior Processes, 23*(4), 491.
Zelazo, P. D., Carter, A., Reznick, J. S., & Frye, D. (1997). Early development of executive function: A problem-solving framework. *Review of general psychology, 1*(2), 198.
Zentall, T. R., Edwards, C. A., Moore, B. S., & Hogan, D. E. (1981). Identity: The basis for both matching and oddity learning in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 7*(1), 70.
Zhang, W., & Luck, S. J. (2008). Discrete fixed-resolution representations in visual working memory. *Nature, 453*, 233–235.