
Volume 34, No. 2 April 2011
ISSN 0163-6383
34 (2) 215–382 (2011)
ELSEVIER
Infant Behavior & Development
An International and Multidisciplinary Journal
EDITOR • Geert Savelsbergh
Vrije Universiteit Amsterdam
The Netherlands

Infant Behavior & Development 34 (2011) 321–326

Contents lists available at ScienceDirect

Infant Behavior and Development

Assessing categorization performance at the individual level: A comparison of Monte Carlo Simulation and Probability Estimate Model procedures

Martha E. Arterberry[a,*], Marc H. Bornstein[b], O. Maurice Haynes[b]

[a] Department of Psychology, Colby College, 5553 Mayflower Hill, Waterville, ME 04901, USA
[b] Child and Family Research, Eunice Kennedy Shriver National Institute of Child Health and Human Development, USA

ARTICLE INFO

Article history:
Received 28 May 2010
Received in revised form
10 September 2010
Accepted 3 February 2011

Keywords:
Categorization
Categorizer classification
Toddlers

ABSTRACT

Two analytical procedures for identifying young children as categorizers, the Monte Carlo Simulation and the Probability Estimate Model, were compared. Using a sequential touching method, children aged 12, 18, 24, and 30 months were given seven object sets representing different levels of categorical classification. From their touching performance, the probability that children were categorizing was then determined independently using Monte Carlo Simulation and the Probability Estimate Model. The two analytical procedures resulted in different percentages of children being classified as categorizers. Results using the Monte Carlo Simulation were more consistent with group-level analyses than results using the Probability Estimate Model. These findings recommend using the Monte Carlo Simulation for determining individual categorizer classification.

1. Introduction

Questions of when and how children organize objects and events in their world into categories have occupied philosophers and psychologists for over half a century (Ricciuti, 1965; Starkey, 1981; see Rakison & Oakes, 2003; Cohen & Cashon, 2006, for reviews). Integral to addressing these questions is the best way to measure existing categories and the development of new categories in children. The sequential touching method is commonly used to assess categorization abilities of children between the ages of 1 and 3 years (e.g., Mandler, Fivush, & Reznick, 1987). In the sequential touching method, children are presented simultaneously with a collection of objects from two categories (e.g., four animals and four vehicles), and their patterns of touching are observed, recorded, and assessed. The empirical observation is that, if children recognize a categorical distinction amongst the objects, they touch those from within one or the other category in succession more than would be expected by chance.

Sequential touching is analyzed at a group level and at an individual level. At the group level, aggregating over individual data, mean run length is the focus of analysis (see Mandler et al., 1987). Run length is the number of touches in a row to objects from the same category. A run can range from 1 (if the child touches only 1 object from 1 category before touching an object from another category) to the total number of the child's touches (if the child touches only objects from one category). For each object category, the mean of all run lengths is calculated. Children as a group touching objects from the same category at run lengths greater than chance leads to the inference that they categorize.

* Corresponding author. Tel.: +1 207 859 5553; fax: +1 207 859 5555.
E-mail addresses: Martha.Arterberry@colby.edu, marterbe@colby.edu (M.E. Arterberry).

0163-6383/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.infbeh.2011.02.003

Mean run length provides key information regarding whether children as a group categorize objects. Mean run length, however, does not indicate whether individual children categorize objects from one or both categories. For example, suppose a child is presented with 4 different dogs and 4 different horses (e.g., Kovack-Lesh & Oakes, 2007). The child might touch three of four horses but only one dog and still produce a high mean run length, showing an overall high level of categorization. To complement analyses at the group level, assessment is also made at the level of individual children.

At the individual level of analysis, the question is whether each individual child categorizes. Traditionally, this question has been addressed by Monte Carlo Simulation (MCS; Mandler et al., 1987). Evaluation is based on the child's total number of touches, the longest run length for a category in an object set, and whether the child touched at least 3 or 4 different exemplars from the same category in the set. Because a run of touching multiple unique objects in a row can occur by chance, especially when a child makes many touches, a Monte Carlo program determines the likelihood of occurrence of runs. The program computes how often categorizing runs occur in 10,000 random draws. Repetitions are allowed (excluding touches to the same object in immediate succession) as long as a run includes 3 or 4 unique objects. This technique estimates the probability of one or more categorizing runs occurring by chance, as a function of the total number of objects a child touches.

The Monte Carlo Simulation for categorizer classification is widely used; however, it has been criticized on at least two grounds (e.g., Thomas & Dahlin, 2000). First is a criticism pertaining to the criterion of touching at least 3 different objects out of 4 in a category, a criterion that does not appear to have a theoretical foundation. Thomas and Dahlin (2000) asked "Why three or four different objects and not two?" (p. 183). A second criticism of Monte Carlo Simulation is its reliance on the longest run to a set of objects. The longest run represents only a subset of the child's touches. Thomas and Dahlin (2000) instead suggested that average run length to objects in the set across the session would better capture an individual child's categorization performance.

To address these criticisms, Thomas and Dahlin (2000) proposed an alternative analytic procedure, referred to here as the Probability Estimate Model (PEM). In PEM, children's mean run length to one of the two categories is assessed in terms of the number of touches of objects in the category and a theoretical distribution of run lengths. This model assumes that a child's sequence of touches follows one probability distribution if the child is a categorizer and another distribution if the child is not a categorizer. Mean run length is computed for each object set separately to obtain a probability estimate for each child having categorized the objects. Thus, on trials in which children are presented with 4 dogs and 4 horses, children receive a mean run length for dogs and a mean run length for horses. In both procedures, children are classified as categorizers for both categories in each set (e.g., animals and vehicles). If they were classified as categorizers for one of the two categories in the set (e.g., animals or vehicles), they were labeled a "single categorizer" for that category. If they were classified as categorizers for both categories in the set (e.g., animals and vehicles), they were labeled "dual categorizers".

The goal of the present analysis was to compare individual categorizer assignments by applying MCS and PEM to a published data set (Bornstein & Arterberry, 2010). To this end, children aged 12, 18, 24, and 30 months were given 7 sets comprised of 8 eight objects each. In each set, half of the objects were from one category (e.g., horses) and half the objects were from a different category (e.g., dogs). The sets compared categorization of animals, vehicles, fruit, and furniture at varying levels of inclusiveness. Each child was then classified as a categorizer or not using the Monte Carlo Simulation and the Probability Estimate Model. We compared the two analytic procedures in terms of percentages of children identified as categorizers and mean run length differences. Analyses of age, domain, and level of inclusiveness were not addressed, as they are reported elsewhere along with results of mean run length differences (Bornstein & Arterberry, 2010). The present analyses were organized around two questions: Do the two analytical procedures, Monte Carlo Simulation and Probability Estimate Model, identify the same percentages of children as single and dual categorizers? Which procedure, Monte Carlo Simulation or Probability Estimate Model, best converges with group level performance, as indexed by mean run length of children identified as categorizers?

2. Method

The method is reported in detail in Bornstein and Arterberry (2010); thus, only a brief over view is provided below.

2.1. Participants

Twenty 12-month-olds (*M* age = 12 months, 7 days, *range* = 12 months, 2 days to 12 months, 14 days), 20 18-month-olds (*M* age = 18 months, 7 days, *range* = 17 months, 24 days to 18 months, 14 days), 20 24-month-olds (*M* age = 24 months, 7 days, *range* = 23 months, 15 days to 24 months, 10 days), and 20 30-month-olds (*M* age = 30 months, 7 days, *range* = 29 months, 14 days to 30 months, 14 days) participated in the study.

2.2. Materials and procedure

Small naturalistic three-dimensional scale models were used to create 14 sets of stimuli (listed in Tables 1 and 2). Each set contained 4 replica objects from 2 categories, animals and vehicles or fruit and furniture.

Table 1
Percentages of children classified as single or dual categorizers tested with animal–vehicle sets using the Monte Carlo Simulation (MCS) and the Probability Estimate Model (PEM) procedures.

	Age															
	12 months				18 months				24 months				30 months			
	MCS		PEM		MCS		PEM		MCS		PEM		MCS		PEM	
	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual
Animal/vehicle	50	0	90	10	70	10	40	0	80	20	100	0	60	30	0	100
Frogs/cows	40	0	10	0	60	10	90	10	40	50	40	0	60	20	50	50
Helicopters/pickups	30	30	20	0	80	0	80	20	30	50	70	30	50	30	70	0
Dogs/horses	30	30	10	0	20	0	90	10	10	40	0	0	60	10	100	0
SUVs/trucks	30	20	10	0	40	10	20	0	60	10	0	100	50	10	10	0
Mako/hammerhead	30	0	0	100	20	30	0	100	30	20	10	0	10	20	0	0
Convertibles/hardtops	30	30	100	0	10	30	0	100	50	10	100	0	10	40	0	100

The child sat at a small table, either on a parent's lap or alone in a chair (in which case the parent sat behind the child). The experimenter sat opposite the child. A tray was placed on the table in front of the child, enabling the simultaneous presentation of all objects. A camera positioned behind the experimenter focused on the child's torso and recorded the child's actions with the objects on each trial.

Children were tested in either an animal–vehicle or fruit–furniture condition. On each trial, the experimenter informally and randomly positioned the 8 objects (4 from each domain) on the tray in front of the child. After placing the tray on the table within easy reach of the child, the experimenter gave the standard prompt, "These are for you to play with." Children were allowed to manipulate the objects in any way they wished for 2 min with no further prompting.

2.3. Scoring

Video records were scored randomly by a single coder who was naïve to the hypotheses of the study. The order in which the child touched objects was coded. A second coder who was also naïve to the hypotheses coded a random sample of 25% of the sessions to obtain a measure of coding reliability for touches. Agreement was based on each object contact, and a value reflecting percent agreement was calculated for each set for each child. Mean agreement was 90% (*range* = 86–94%).

From this scoring, the order of objects and categories sequentially touched were derived to calculate mean run length. To assess individual level the Monte Carlo Simulation and the Probability Estimate Model were both used. Categorizer classification based on the Monte Carlo used the total number of touches to objects in the category, the longest run to objects in that category, and whether the child touched 3 or 4 of the objects in the category. The probability associated with these criteria occurring by chance was determined using the look up table provided by Dixon, Woodard, and Merry (1998). Children with probabilities < .10 were classified as categorizers (see Mandler et al., 1987). For the Probability Estimate Model, the algorithm derived by Thomas and Dahlin (2000) was implemented in R (R Core Development Team, 2005). Categorizer classification was based on mean run length for each category by dividing the number of touches by the number of runs to objects in the category for each age group separately (e.g., 12-month-old animals, 12-month-old vehicles, 12-month-old frogs, etc.). Children were assigned a probability value regarding whether they were a categorizer. Children given a probability estimate of .90 or above were considered categorizers (a criterion consistent with the Monte Carlo Simulation).

Table 2
Percentages of children classified as single or dual categorizers tested with fruit–furniture sets using the Monte Carlo Simulation (MCS) and the Probability Estimate Model (PEM) procedures.

	Age															
	12 months				18 months				24 months				30 months			
	MCS		PEM		MCS		PEM		MCS		PEM		MCS		PEM	
	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual	Single	Dual
Fruit/furniture	60	0	60	0	70	0	70	30	50	10	40	0	30	10	40	0
Grapes/bananas	60	30	60	40	70	10	100	0	50	20	90	10	70	20	10	90
Beds/tables	50	0	100	0	80	0	0	0	40	30	100	0	40	20	0	0
Apricots/lemons	30	0	10	0	20	30	10	10	40	30	90	0	40	20	10	0
Sofas/chairs	20	0	90	10	20	0	90	10	10	0	0	100	30	10	100	0
Red delicious/winesap apples	40	0	50	0	50	0	20	0	40	20	100	0	20	0	90	10
Straight chair/high chair	30	0	90	0	60	0	10	0	20	10	0	0	60	20	100	0

Table 3
Comparisons of mean run lengths to chance (1.75) for categorizers (single or dual) as determined by the Monte Carlo Simulation and the Probability Estimate Model procedures for each animal–vehicle set collapsed across age.

	MCS				PEM			
	Number of categorizers	*M*	*SD*	*t*	Number of categorizers	*M*	*SD*	*t*
Animal	22	3.22	2.05	3.50**	24	2.67	2.12	2.12*
Vehicle	16	4.68	4.98	2.35*	21	3.81	4.57	2.07+
Frogs	23	3.83	1.91	5.22**	11	4.93	2.21	4.78**
Cows	13	2.99	.92	4.87**	20	2.45	1.14	1.96+
Helicopters	24	3.05	.97	6.58**	26	2.80	1.15	4.62**
Pickups	17	4.78	2.30	4.71**	8	6.17	2.13	5.87**
Dogs	14	2.63	.94	3.51**	11	1.65	.87	.39
Horses	14	2.65	.78	4.33**	11	2.64	1.11	2.66*
SUV	13	2.63	1.69	1.87+	11	1.91	.54	.98
Semitruck	15	3.60	2.87	2.49*	13	2.99	3.15	1.42
Mako	13	3.24	1.86	2.89*	21	2.38	1.77	1.63
Hammerhead	10	2.54	.65	3.86**	19	1.66	.73	.50
Convertibles	17	2.65	1.04	3.55**	20	2.08	1.12	1.33
Hard tops	15	2.10	.52	2.60*	40	1.64	.63	1.13

For two-tailed tests: $+p < .10$, $^{*}p < .05$, $^{**}p < .01$.

3. Results

3.1. Monte Carlo Simulation versus Probability Estimate Model

Tables 1 and 2 present the percentages of single and dual categorizers using the Monte Carlo Simulation and the Probability Estimate Model procedures for animal–vehicle and fruit–furniture sets, respectively. Inspection of these tables reveals that the two approaches do not converge very well. For example, consider the percentage of 12-month-old categorizers of convertibles and hard top sports cars (Table 1). Using the Monte Carlo Simulation, 30% were classified as single categorizers and 30% were classified as dual categorizers. Using the Probability Estimate Model, 100% were classified as single categorizers and none was classified as a dual categorizer. Correlation coefficients calculated for percentages of children classified as categorizers using the Monte Carlo Simulation and the Probability Estimate Model showed no significant relation for either single and dual categorizers, $r(55) = .07$, and $r(55) = -.03$, respectively.

3.2. Categorizer classification and mean run length

Researchers using the sequential touching procedure typically use mean run length to evaluate group performance (e.g., Ellis & Oakes, 2006; Mandler, Bauer, & McDonough, 1991; Rakison & Butterworth, 1998). For children classified as categorizers using the Monte Carlo Simulation and the Probability Estimate Model, we tested their mean run length for each category versus chance (1.75, see Mandler et al., 1987). We expected all of the mean run lengths to significantly exceed chance because the only children included in the analysis were categorizers. The results are shown in Tables 3 and 4. When using the Monte Carlo Simulation, all 14 comparisons for animal–vehicle sets and 12 comparisons for fruit–furniture sets exceeded chance. When using the Probability Estimate Model, 7 of the animal–vehicle comparisons and 3 of the fruit–furniture comparisons exceeded chance.

4. Discussion

The purpose of this investigation was to evaluate the relative usefulness of the Monte Carlo Simulation and the Probability Estimate Model in classifying individual young children as categorizers in a sequential touching paradigm. To this end, we first compared percentages of children classified as categorizers using the Monte Carlo Simulation to the percentages obtained using the Probability Estimate Model to address the first question, specifically do the analytical procedures identify the same percentages of children as single and dual categorizers? The two approaches yielded different percentages, and there was little correlation between the two for single and dual categorizers. The two approaches rely on different assumptions and analyses. The Monte Carlo Simulation requires that a child touch 3 of 4 objects in a set and have a longest run that is significantly greater than would be expected by chance (e.g., Mandler et al., 1987). The Probability Estimate Model determines the probability that a child is a categorizer based on his/her mean run length to the items in a category (e.g., dogs only when presented with a set containing dogs and horses). Thus, a child's mean run length is evaluated in the context of other mean run lengths by children of the same age to the same category (Thomas & Dahlin, 2000). Given the fact that the analytic

Table 4
Comparisons of mean run lengths to chance (1.75) for categorizers (single or dual) as determined by the Monte Carlo Simulation and the Probability Estimate Model procedures for each fruit–furniture set collapsed across age.

Fruit–furniture	MCS				PEM			
	Number of categorizers	*M*	*SD*	*t*	Number of categorizers	*M*	*SD*	*t*
Fruit	16	4.27	2.22	4.53**	12	4.72	2.47	4.15**
Furniture	9	4.11	1.88	3.76**	15	2.13	1.92	.77
Grapes	30	4.42	3.25	4.50**	35	4.11	3.12	4.47**
Bananas	11	2.99	1.59	2.58*	19	2.35	1.43	1.84
Beds	18	2.76	1.49	2.88*	20	1.87	.75	.69
Tables	13	2.51	1.02	2.68*	0	–	–	–
Apricots	13	3.34	1.78	3.20**	0	–	–	–
Lemons	16	4.15	3.74	2.57*	13	3.79	4.26	1.73
Sofas	5	2.28	.74	1.60	31	1.58	.45	−2.11*a
Chairs	5	2.19	.47	2.01	21	1.79	1.26	.14
Red delicious	10	3.84	3.59	1.84+	3	7.33	5.51	1.76
Winesaps	9	2.67	.61	4.49**	25	1.62	.70	.91
Straight chairs	12	2.44	.26	9.04**	10	1.45	.48	1.97+
High chairs	11	2.51	.82	3.09*	10	1.91	.60	.83

For two-tailed tests: *a* significantly below chance +$p < .10$, *$p < .05$, **$p < .01$.

procedures use different criteria and dependent measures, it is not surprising that the percentages of children classified as single and dual categorizers were not consistent across the two analytic procedures.

The second question addressed which procedure more closely articulates with group-level performance. To this end, the mean run lengths for categorizers identified by the two analytical procedures were compared to chance. For 26 out of 28 comparisons, mean run lengths of children classified as categorizers using the Monte Carlo Simulation were consistently above chance, as would be expected because all these children were identified as categorizers. For the Probability Estimate Model, less than one half of the comparisons resulted in mean run lengths exceeding chance. These results favor the Monte Carlo Simulation for individual categorizer classification.

It should be noted that both analytic procedures resulted in an attempt to classify individual children as categorizers. We did not, however, have independent verification that these children were indeed categorizing the objects in each set. Children classified as categorizers based on the Monte Carlo Simulation as a group showed mean run lengths significantly greater than chance, but without another individual measure of categorization, we can only make a probabilistic judgment regarding whether any individual child was a categorizer.

The ability to estimate whether individual children are categorizers complements group-level analysis that typically uses mean run length. The combination of group-level and individual-level analyses provides a more complete picture of the abilities of young children. It is not uncommon to find that a group of children as a whole shows poor categorization performance, yet a subset of the children in that group still appear to be categorizers based on individual analyses (e.g., Bornstein & Arterberry, 2010). As is well known from introductory psychology and introductory political science, the group does not dictate the individual.

Acknowledgements

We thank T. Taylor and H. Thomas for assistance. This research was supported by the Intramural Research Program of the NIH, NICHD.

References

Bornstein, M. H., & Arterberry, M. E. (2010). The development of object categorization in young children: Hierarchical inclusiveness, age, perceptual attribute and group versus individual analyses. *Developmental Psychology*, *46*, 350–365. doi:10.1037/a0018411

Cohen, L. B., & Cashon, C. H. (2006). Infant cognition. In W. Damon, D. Kuhn, & R. Siegler (Eds.), *The handbook of child psychology: Cognition, perception, and language* (pp. 214–251). Hoboken, NJ: John Wiley & Sons, Inc.

Dixon, W. E., Woodard, T., & Merry, M. S. (1998). TouchStat: A Monte Carlo program for calculating sequential touching probabilities. *Behavior Research Methods, Instruments, & Computers*, *30*, 592–604.

Ellis, A. E., & Oakes, L. M. (2006). Infants flexibly use different dimensions to categorize objects. *Developmental Psychology*, *42*, 1000–1011. doi:10.1037/0012-1649.42.6.1000

Kovack-Lesh, K. A., & Oakes, L. M. (2007). Hold your horses: How exposure to different items influences infant categorization. *Journal of Experimental Child Psychology*, *98*, 69–93. doi:10.1016/j.jecp.2007.05.001

Mandler, J. M., Fivush, R., & Reznick, J. S. (1987). The development of contextual categories. *Cognitive Development*, *2*, 339–354. doi:10.1016/S0885-2014(87)80012-6

Mandler, J. M., Bauer, P. J., & McDonough, L. (1991). Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology*, *23*, 263–298. doi:10.1016/0010-0285(91)90011-C

Rakison, D. H., & Butterworth, G. E. (1998). Infants' use of object parts in early categorization. *Developmental Psychology*, *34*, 49–62. doi:10.1037/0012-1649.34.1.49

Rakison, D. H., & Oakes, L. M. (2003). *Early category and concept development: Making sense of the blooming buzzing confusion*. Oxford, UK: Oxford University Press.

R Core Development Team. (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Ricciuti, H. (1965). Object grouping and selective ordering behavior in infants 12 to 24 months old. *Merrill-Palmer Quarterly*, *11*, 129–148.

Starkey, D. (1981). The origins of concept formation: Object sorting and object preference in early infancy. *Child Development*, *52*, 489–497. doi:10.2307/1129166

Thomas, H., & Dahlin, M. P. (2000). Inferring children's categorizations from sequential touching behaviors: An analytical model. *Psychological Review*, *107*, 182–194. doi:10.1037/0033-295X.107.1.182