The Development of Clustering in Episodic Memory: A Cognitive-Modeling Approach

Sebastian S. Horn

University of Zürich

Ute J. Bayen and Martha Michalkiewicz

Heinrich-Heine-Universität Düsseldorf

Author Note

Sebastian S. Horn, Department of Psychology, University of Zürich, Switzerland; Ute Johanna Bayen and Martha Michalkiewicz, Institute for Experimental Psychology, Heinrich-Heine-Universität Düsseldorf, Germany.

The authors thank Marcus Haseloff for the voice recordings, Siegmund Switala, Fabian Gümüsdagli, and Dennis Rohrschneider for help with technical aspects of the project, Alina Noll, Anastasia Wagner, Christina Engels, Daniela Czernochowski, and Sarah Schnitzler for help with data collection, and Daniel Heck and Morten Moshagen for helpful comments regarding the analyses. Online supplemental materials and open data for this study are available at *https://tinyurl.com/devpsychrecall*.

Correspondence concerning this article should be addressed to Sebastian S. Horn, University of Zürich, Department of Psychology, Binzmühlestr. 14 Box 11, 8050 Zürich, Switzerland, Email: horn@psychologie.uzh.ch, or to Ute J. Bayen, Heinrich-Heine-Universität Düsseldorf, Institute for Experimental Psychology, Geb. 23.02., Universitätsstr. 1, 40225 Düsseldorf, Germany, Email: ubayen@hhu.de

Abstract

Younger children's free recall from episodic memory is typically less organized than recall by older children. To investigate if and how repeated learning opportunities help children use organizational strategies that improve recall, the authors analyzed category clustering across four study-test cycles. 7-year-olds, 10-year-olds, and young adults (*N*=150) studied categorically related words for a free-recall task. The cognitive processes underlying recall and clustering were measured with a multinomial model. The modeling revealed that developmental differences emerged particularly in the rate of learning to encode words as categorical clusters. The learning curves showed a common pattern across age groups, indicating developmental invariance. Memory for individual items also contributed to developmental differences and was the only factor driving 7-year-olds' moderate improvements in recall.

Keywords: cognitive development, episodic memory, free recall, category clustering, multinomial processing tree modeling, Bayesian hierarchical modeling

Data and Online Supplemental Materials: https://tinyurl.com/devpsychrecall

The Development of Clustering in Episodic Memory: A Cognitive-Modeling Approach

Episodic memory refers to the ability to remember specific events from one's personal past. It supports daily functioning (remembering what, when, or where something was experienced) and contributes to building an identity over time (Schacter & Tulving, 1994). A key factor contributing to episodic-memory development from childhood to young adulthood is the ability to form connections between objects, events, or persons. In memory tasks that involve semantically related items, for instance, adults and older children often organize their responses by category and recall related items adjacently—a phenomenon known as *category clustering*. In contrast, younger children (below ~8 years of age) often recall related items without such clustering; they may already understand the meaning of a given item (e.g., "a spoon is used for eating"), but there is typically a lag of several years until children connect this meaning across multiple exemplars of a category (e.g., "spoons, forks, knives, are all types of cutlery"). Overall, a wealth of developmental research has found that younger children's memory output is semantically less organized than that of older children and adults and that the spontaneous formation of relations among items increases with age (e.g., Bjorklund & Jacobs, 1985; Hasselhorn, 1990). Major changes in category clustering appear to emerge around ages 6 to 12—a phase in which children's knowledge base also expands substantially through education and experience (Schneider, 2014). As category clustering is associated with good recall performance (Bjorklund, 2011), it is important to understand how clustering develops across childhood and how it can be facilitated.

In the present study, we used cognitive modeling to investigate developmental differences in the processes underlying clustering in free recall. One important approach toward understanding the organizational principles of episodic memory rests on the comparison of originally presented information and the structure of people's responses (Bower, 1970; Mandler, 1967). Starting in the 1960s, much research has examined the influence of different presentation formats, materials (Cole, Frankel, & Sharp, 1971), rehearsal strategies (Ornstein, Naus, & Liberty, 1975), and organizational strategy instructions (e.g., Moely, Olson, Halwes, & Flavell, 1969; Rao & Moely, 1989) on children's free recall. Analyses that involved only one single learning opportunity have frequently been used to assess whether people of different ages notice and strategically utilize relational information for recall (for overviews, see Jablonksi, 1974; Schneider, 2014). Little is known, however, about the cognitive dynamics of categorical clustering when repeated learning opportunities are available. In what follows, we discuss this aspect and consider the few developmental studies that focused on changes in categorical clustering in multitrial free recall.

Repeated Learning Opportunities

Recall tasks that include multiple study-test trials make it possible to measure progressive changes in memory organization over time and to examine whether children and adults access and successively reorganize information in different ways. Many diagnostic memory tests use repeated study-test trials because individual differences in performance are often most pronounced in the rate of learning across trials (e.g., Bröder, Herwig, Teipel, & Fast, 2008). Moreover, as noted by Paris (1978), learning activities in educational settings often involve cyclical or recursive memorizing. Proficiency in deliberate recall is often relevant at school, but is unlikely familiar to preschool children. Thus, a lack of experience with tasks that require memorization could be one important factor explaining why clustering strategies are hardly observed in younger children (Glidden, 1977). If younger children recall fewer items because they are unfamiliar with effective strategies for recall, then repeated exposure could lead to adaptive changes in the way they utilize presented information. In other words, younger children might need more time or opportunities than older children to notice conceptual relations between items and the categorical structure of a word list. Therefore, one possibility is that younger children initially show complete absence or low levels of categorical clustering in free recall, but increasingly employ such strategies when learning opportunities are granted more frequently. Repeated exposure to the same items in free recall could have similar effects as other manipulations that are well-known to stimulate the production of mnemonic strategies—even in younger children (e.g., organizational instructions; Moely et al., 1969). One aim of our study was to test this possibility.

Regarding changes in clustering over trials, experimental findings on children's free recall are equivocal. For example, Glidden (1977) found age differences in recall performance between 6-year-olds and 9-year-olds. However, increases in subjective organization of items over training sessions were similar in these age groups (the items in Glidden's study, however, were categorically unrelated). Using lists with categorically related item pairs (drawings of common objects), Moely and Shapiro (1971) also reported age differences in recall and clustering in four groups of children (3, 4, 5, and 6-7 years). Increases in clustering emerged across sessions when related items were consistently presented together over trials, but there were no interactions involving age. These findings suggest that children of different ages benefit in similar ways from repeated learning opportunities. In contrast, Cole et al. (1971) consistently found Age × Trial interactions in three experiments with school-age children from grades 1 to 9. Free recall performance and clustering increased across study-test trials in all age groups; however, the changes were substantially smaller in younger than older children, implying that age differences in performance increased with training. Taken together, multiple trials seem to improve free recall as well as clustering in school-age children. It seems unclear, however, whether (a) older children benefit more than younger children from repeated learning opportunities and (b) which cognitive processes might underlie such differences.

In the current research, we examined whether age differences in category clustering over multiple presentations of related items are attributable to children's learning to encode relational information more effectively. This may involve increasing proficiency in simultaneously attending to multiple pieces of information (Halford, Halford, Wilson, & Phillips, 1998) and in forming connections among them, based on semantic knowledge (Bjorklund, 2011; Schneider, 2014). The perspective that the development of encoding processes is an important source of age differences in episodic memory is not uncontested, however, as there is also evidence that the ability to retrieve episodic information may be particularly fragile during early development (Bauer, Wiebe, Carver, Waters, & Nelson, 2003) and that younger children's retrieval is more rigidly dependent on contextual cues from an original encoding phase (e.g., Ackerman, 1981). Addressing these issues is important for our understanding of the development of episodic memory. To investigate our assumptions about developmental differences in the encoding of items as clusters, it is essential to measure cluster encoding separately from other cognitive processes, namely, the retrieval of clusters once they are encoded, and memory for (unclustered) individual items (which may reflect non-strategic, rote memorization). To obtain unconfounded measures of these cognitive processes, we used a modeling approach. We will next discuss advantages of cognitive modeling and then describe the specific model we used in the current study.

Measuring Clustering in Episodic Memory

Various behavioral measures of clustering and subjective organization have been developed and commonly applied to quantify input-output relations in free recall (Bousfield & Bousfield, 1966). An example of such a behavioral measure is the number of clusterable items recalled adjacently. There are several potential reasons for disparate findings about clustering in developmental memory research. First, behavioral measures are not process pure: Observable performance results from a conglomerate of different underlying cognitive components (e.g., cluster encoding and retrieval) that may even interact in a nonlinear fashion (e.g., Brainerd, 1985). For example, items that are not organized in clusters at recall may indicate a failure of cluster encoding or a failure to retrieve an encoded cluster. Second, behavioral measures often pose difficulties for interpretation because they are not tied to a theoretically motivated model and make implicit or untested assumptions (Jablonksi, 1974; Riefer, 1982). Therefore, an important advantage of cognitive modeling over the use of behavioral measures is that the assumptions of formalized cognitive models are explicit and statistically testable; moreover, quantitative estimates of cognitive constructs are obtained in a well-specified way. For instance, Howe, Brainerd, and Kingma (1985) used a two-stage Markov model to examine the impact of the organizational structure of item lists on encoding, retrieval, and retrieval learning in 7-year-olds and 11-year-olds. Their study revealed that the probability of encoding and retrieving information was higher in older than younger children. However, there was surprisingly little increase through presence of semantic list structures in any of the model parameters. This may suggest that the presentation of organized information at encoding is not sufficient to boost memory performance. However, Howe et al.'s Markov model does not provide estimates of item clustering. For the present study, we needed a model that allowed us to estimate the probability to encode item pairs as clusters, unconfounded by other cognitive processes. To this end, we used the pair-clustering model introduced by Batchelder and Riefer (1980, 1986), which is a stochastic model belonging to the class of multinomial processing tree (MPT) models (for overviews of MPT modeling, see Batchelder & Riefer, 1999; Erdfelder et al., 2009).

MPT Modeling

MPT modeling is a powerful tool to disentangle different cognitive states or processes that jointly contribute to observable responses in specific cognitive tasks. MPT models have been successfully applied in many fields of cognitive science, including the development of memory and decision making (e.g., Bender, Wallsten, & Ornstein, 1996; Bernstein, Erdfelder, Meltzoff, Peria, & Loftus, 2011; Chechile, Richman, Topinka, & Ehrensbeck, 1981; Horn, Ruggeri, & Pachur, 2016; Howe, & O'Sullivan, 1997; Pohl, Bayen, & Martin, 2010; Yim, Dennis, & Sloutsky, 2013).

The pair-clustering model. Figure 1 shows the tree structure of the pair-clustering MPT model (Batchelder & Riefer, 1980, 1986) which is tailored to episodic free recall tasks in which pairs of items that belong to the same category (e.g., *spoon* and *knife*) are presented during a study phase. The pair-clustering model provides probability estimates of entering specific cognitive states that can be estimated from categorical response data: Parameter c indicates the probability of encoding related items together as a cluster and storing this cluster over time; parameter r indicates the probability of subsequently retrieving such a cluster at test; parameter u indicates the probability of recalling items that are encoded and retrieved without such clustering (i.e., remembering an item as a singleton without forming a cluster with the other item of the same category). The model distinguishes four mutually exclusive and exhaustive response events (E_i) in free recall: both items of a pair are recalled adjacently (E_1); both items of a pair are recalled, but not adjacently (E_2); only one item of a pair is recalled and the other one not (E_3); neither item of a pair is recalled (E_4). Hence, the model accounts not only for the number of items recalled but also for the structure of the output and makes some basic assumptions about how latent cognitive states lead from a study list to the observed responses: With probability *c*, the items of a pair are encoded and stored together as a clustered representation. In a subsequent recall test, such a cluster may be retrieved (with probability *r*) or not retrieved (with probability 1 – *r*). If the cluster is retrieved, participants recall both items adjacently (E_1). If the cluster is not retrieved, participants do not recall either item of the pair (E_4). Items of a pair that have not been encoded as a cluster (with probability $1 - c$) may nevertheless be encoded and later retrieved as unclustered singletons, each with probability u. With probability $u \times u$, both items of a pair are recalled as singletons, resulting in non-adjacent recall of both words (E_2). With probability u × (1 – *u*), only one of the items is recalled as a singleton (E_3). Finally, with probability (1 – *u*) × (1 – *u*), neither item is recalled as a singleton. In this case, participants do not recall either item of the pair (E_4).

The model has been applied to many data sets (including recall data from younger and older adults; Bröder et al., 2008; Erdfelder & Bayen, 1991; Riefer & Batchelder, 1991) and has shown good fit to recall data. Tests of selective influence of experimental manipulations on the model parameters supported the validity of the psychological interpretation of the parameters (e.g., Bäuml, 1991; Riefer et al., 2002).

Multitrial extension of the model. As developmental differences may emerge in the rate of learning to cluster, which we sought to quantify, we used an extension of the pair-clustering model that accounts for changes in the probabilities of entering cognitive states across multiple trials. Knapp and Batchelder (2004) proposed a framework for reparametrizing MPT models under the constraint of a weak ordering of parameter values. In multitrial learning designs, this entails the plausible assumption that parameter values do not decrease across trials (i.e., the values of *c*, *r*, and *u* may increase across trials or remain at the same level). Specifically, the reparameterized model yields an estimate of parameter θ_1 for the initial study-test trial and a corresponding change-rate parameter $\beta_{\theta,n}$ which models the change of parameter θ from trial $n - 1$ to trial n. This change is quantified as the proportional reduction of error in the modeling: $\beta_{\theta,n} = \frac{\theta_n - \theta_{n-1}}{1-\theta_{n-1}}$. For instance, if the value of some parameter in the initial trial was $\theta_1 = .10$ and on the subsequent trial $\theta_2 = .20$, then the change rate between these trials would be quantified as $\beta_{\theta,2} = \frac{.20-.10}{1-.10} \approx .11$. Further mathematical details are discussed in Knapp and Batchelder (2004). One restricted version of this multitrial model assumes a single change rate across trials that remains constant $\beta_{\theta,2} = \beta_{\theta,3} = \cdots = \beta_{\theta,n}$ for each parameter θ and thus predicts geometric/exponential learning curves (Bush & Mosteller, 1955). In the current study, we applied this restricted model and also compared it with a more flexible model version that allows any form of monotonic increase across study-test trials.

Hierarchical Bayesian implementation. We used a hierarchical MPT model (Klauer, 2010; Matzke, Dolan, Batchelder, & Wagenmakers, 2015) which can account for individual differences in the cognitive components underlying free recall. As heterogeneity in performance and underlying cognitive processes is expected particularly in samples of children, this is a notable improvement over previous MPT analyses in developmental studies, for which data were aggregated over participants (e.g., Smith, Bayen, & Martin, 2010). Because hierarchical implementations of MPT models in a classic-frequentist framework are sometimes difficult (Klauer, 2010), the current parameter estimation relied on a Bayesian approach in which uncertainty about model parameters and available information is represented by probability distributions. Specifically, to determine the most credible value ranges of the model parameters in the posterior distributions given the data, we used the Markov chain Monte Carlo (MCMC) methodology for posterior sampling (Van Ravenzwaaij, Cassey, & Brown, 2018, provided an introduction; further details about the current model implementations are in the Appendix and in the Online Supplement).

The Current Study

The objective of the cross-sectional study reported here was to investigate if and how the formation of clusters in free recall and the ability to benefit therein from multiple learning opportunities develops in childhood. 7-year-olds, 10-year-olds, and young adults studied a list of 18 categorically related word pairs that were presented auditorily. We selected these age groups because considerable changes in both clustering and episodic recall performance are expected during the elementary-school years (Bjorklund, 2011; Fandakova, Lindenberger, & Shing, 2015; Ghetti & Lee, 2011; Lehmann & Hasselhorn, 2010; Schneider, 2014). There were four consecutive study-test trials of the same item list. Each study phase was followed by free recall. We used the MPT pair-clustering model to estimate cluster encoding and changes in this process in the three age groups.

Quantitative Comparisons

Initial baseline level. Based on previous developmental research (Jablonski, 1974; Schneider, 2014), we expected differences between younger children (7-year-olds), older children (10-year-olds), and adults, in free-recall performance (recall of words and of word pairs). We also anticipated age differences in the corresponding model-based measures on the initial study-test trial: To the extent that adults and older children more likely engage in spontaneous cluster encoding than younger children, we expected that age differences in recall would be attributable to the initial level of cluster encoding (parameter c). This would be in line with the notion that the proficiency to form semantic relations between items undergoes protracted development during the elementary-school years (e.g., Schneider, 2014). However, strategic cluster encoding may not be the only source contributing to developmental differences in recall. To the extent that age groups differ in remembering individual (unclustered) items, we may also anticipate differences in parameter u.

Learning rate. Developmental differences may be more pronounced in the way people progressively utilize information than in their initial performance. Therefore, we were particularly interested in changes of cognitive processes across study-test trials. As discussed above, one possibility is that repeated learning opportunities help younger and older children improve clustering and recall to similar extent (Scenario A). A second possibility is that younger children's lack of experience with clustering strategies may be compensated for by repeated learning opportunities. In this case, we would find greater benefit from repetitions for younger children's clustering than for older children's clustering (Scenario B). Both of these result patterns would indicate that even younger children are able to produce clustering strategies if given such training opportunities (Moely & Shapiro, 1971). A third possibility (Scenario C) is that older children's broader knowledge base, or a failure of younger children to overcome a production deficiency result in older children benefitting more from multiple learning opportunities than younger children (Cole et al., 1971). In the modeling, these three possibilities would be reflected in different patterns of change rates in cluster encoding (parameter β_c). We aimed at investigating which of these three scenarios best describes children's changes in strategy use when multiple learning opportunities are available. A schematic illustration of the three different predictions is in the Appendix.

Qualitative Comparisons

The cognitive components contributing to recall may also differ qualitatively between age groups. That is, even the same level of memory performance may in principle be achieved through different cognitive routes. Specifically, recall of items may be achieved by memorizing them as singletons in a rote manner or by clustering them by semantic category. Younger children sometimes show excellent memory for individual items (Sloutsky & Fisher, 2004). Moreover, there is evidence to suggest that the basic ability for episodic recall emerges in infancy (e.g., Bauer et al., 2003), whereas the ability to form semantic relations between items undergoes protracted development (Schneider, 2014). Therefore, we assumed that age differences should be more pronounced in categorical clustering than in recall of singletons. The design of the current study along with the use of the cognitive model allowed us to disentangle these possibilities through age comparisons of the model parameters *c* (cluster encoding) and *u* (singleton recall).

Moreover, we analyzed whether the pattern of learning to cluster differed between age groups or whether there was *structural developmental invariance* in multitrial free-recall learning. Developmental invariances refer to regularities in data that hold regardless of age group. The modeling of invariances is theoretically informative because it points to general characteristics or principles of human cognition (see Brainerd, 1983; Lee, 2018). In many cognitive tasks, for example, learning is well described by a constant rate of change in performance (Bush & Mosteller, 1955). If it was possible in the current study to describe the general form of learning with the same functions in all age groups, this would indicate structural developmental invariance. Therefore, we tested if a learning model with constant change rates in model parameters over study-test trials fitted the data from all age groups well, and if age differences were merely confined to the values of the change-rate parameter.

Method

Participants

The study included 150 participants from three age groups: forty-nine 7-year-olds, forty-nine 10-year-olds, and fifty-two young adults. Given these sample sizes and an alpha level of .05, the statistical power to detect medium-sized ($\eta^2 = .06$) main effects or interactions in behavioral recall measures was at least .80 (Faul, Erdfelder, Lang, & Buchner, 2007). Table 1 shows participant characteristics and mean scores in three cognitive tests of the Wechsler intelligence scales. We recruited the 7-year-old children from two first-grade classes and the 10-year-old children from two fourth-grade classes of the same public elementary school in the city Düsseldorf (Germany) through letters of invitation to parents. We obtained approval from the ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich-Heine-University Düsseldorf, written permission from the school administration, and consent and demographic information about the children from their parents. The children provided written assent and received a small toy for participation. Young adults were either first-semester psychology students who received course credit, or students with other majors who received financial reimbursement. Participants had various socioeconomic backgrounds (predominantly middle- and upper-class Caucasian families). Only native speakers of German were included in the study. All participants were screened for language comprehension at the beginning of each session with a short picture-naming task (Bates et al., 2003). Participants showed typical cognitive performance for their respective age (participant scores of the cognitive tests similarities, digit span, and digit-symbol coding are in Table 1).

Design and Materials

We used a 3 × 4 mixed factorial design with the between-subjects factor *age group* (7-year-olds, 10-year-olds, young adults) and the within-subject factor *study-test trial* (trials 1 to 4). The materials and procedure are illustrated in Figure 2. The same word list was presented in each of the four study-test trials. The list included 18 semantically related word pairs (i.e., 36 words) from 18 semantic categories, taken from German category-production norms (Mannhaupt, 1983). The words are listed in Online Supplement 5. All words were one- or two-syllable nouns. We did not use the two most frequently produced exemplars from the production norms to avoid retrieval from semantic memory; however, the selected words had relatively high production frequencies in the norms to ensure high association with a category. The stimuli for the tasks were recorded in advance by a male professional speaker to avoid confounds by reading ability. As primacy buffer, the same 2 word pairs were presented at the beginning of each list in random sequence with a lag of one between the words of a pair and did not enter analyses. Following the procedures by Bröder et al. (2008), for each participant, a random half of the remaining 16 word pairs appeared in the list with a lag of one nonrelated word between them, whereas the other half of word pairs appeared with a lag of nine other words between them. In each study trial, participants heard a different random sequence of the same words, with the restriction that the lag remained constant for a given word pair across study trials. We ensured that selected words were not highly associated with any words from other selected semantic categories by evaluating their co-occurrences (using a German text corpora collection: http://wortschatz.uni-leipzig.de). Moreover, the materials were examined by two independent experts (experienced elementary school teachers) to ensure that first graders would be familiar with all presented words and would understand the instructions. A pilot study with six first-graders indicated that they could successfully handle all instructions, tasks, and materials.

Procedure

Each participant was tested individually in a quiet school room (children) or in a university laboratory (adults) in a session that lasted about 35 min. All participants gave written consent (adults) or assent (children). A trained experimenter explained the tasks and gave standardized oral instructions. After a brief verbal screening test, the participant put on a headset that was connected to a laptop that controlled the auditory presentation of stimuli during the encoding phase and recorded the participant's oral responses during the recall phase. The word stimuli had been pre-recorded by a professional speaker. Participants listened to all stimuli via the headset and responded orally; no interactions with a screen, keyboard, or papers were required. An initial audibility-check, in which participants repeated short sequences of digits, ensured that they could hear the presented sounds via the headset.

The experimenter then informed the participants that they would study a list of words and recall them in any order they wished. The same words would then be repeated several times, but in a different sequence, and each time they would have to recall the words. Participants were instructed to memorize as many words as possible and were informed that they should recall these words after a short delay. Participants had the opportunity to ask the experimenter any questions and were asked to repeat the instructions in their own words to ensure understanding; if necessary this procedure was repeated. For an illustration of the recall procedure, refer to Figure 2. There were four study-test trials. Each of them included the auditory presentation (via headset) of the same 18 semantically related word pairs. Each word was presented for 1s, followed by an inter-stimulus interval of 1s. After list presentation, participants repeated sequences of 3 digits (children) or 6 digits (adults) for 20s (presented by the same voice). This buffer-clearing task served to counteract recall of words from short-term memory.

Participants were then asked to recall as many words as possible in any order. In trials 2-4, participants were additionally informed to also include words that they had already mentioned in the preceding trials. Recording of verbal recall was stopped (a) if participants indicated that they could not remember any more words, or (b) if 20s elapsed without any mentioned item and, following a prompt by the experimenter, no further items were mentioned within the subsequent 20s. At the end of the session, participants completed several cognitive tests (similarities, digit-symbol coding, and forward and backward digit span from the Wechsler intelligence tests). Finally, participants received their reward for participation. Further Verbatim instructions are in Online Supplement 6.

Results

We first report behavioral measures of recall performance, followed by the model-based analyses of cognitive components contributing to free recall and to category clustering.

Recall Performance

Free recall of words. Table 2 shows means and standard deviations of the proportion of words recalled. A 3 (Age Group) × 4 (Trial) ANOVA indicated a large main effect of Age, $F(2, 147) = 159.36, p < .01, \eta^2 = .68$. 7-year-olds recalled fewer words than 10-year-olds, who in turn recalled fewer words than young adults (all *t*s > 6; *p*s < .01; *d*s > 0.5). The large main effect of Trial, $F(3, 441) = 468.98, p < .01, \eta^2 = .62$, Greenhouse-Geisser $\varepsilon = .788$ indicated substantial improvement across trials. Improvement was found within each age group when analyzed separately (all *F*s > 26; all *p*s < .01). However, as indicated by a Group × Trial interaction, the improvement differed between age groups, $F(6, 441) = 72.24, p < .01, \eta^2 = .19$. The age differences in recall were considerably larger in the last study-test trial ($\eta^2 = .74$) than in the initial trial ($\eta^2 = .35$).

Free recall of word pairs. We also examined the proportion of times both words of a pair were recalled (see Table 2). A 3 × 4 ANOVA indicated a large main effect of Age, $F(2, 147) = 139.5, p < .01, \eta^2 = .66$. 7-year-olds recalled fewer pairs than 10-year-olds, who in turn recalled fewer pairs than adults (all *t*s > 4.48; *p*s < .01; *d*s > 0.36). There was a large effect of Trial, $F(3, 441) = 287.68, p < .01, \eta^2 = .49, \varepsilon = .843$. Increases in the recall of pairs across the four trials emerged within each age group (all *F*s > 10; all *p*s < .01). A Group × Trial interaction indicated that improvements across trials in recall of word pairs were more pronounced in older than younger age groups, $F(6, 441) = 77.64, p < .01, \eta^2 = .26$.

Summary. All age groups showed increases across study-test trials in the recall of (a) words and (b) pairs of categorically related words. As expected, there were large differences in overall recall between 7-year-olds, 10-year-olds, and adults. Moreover, improvements through list repetitions were more pronounced with increasing age. However, based on analyses of overall memory performance alone, it is difficult to evaluate whether these age differences are due to differences in cluster encoding, and whether they are mainly attributable to quantitative or qualitative differences in processes. We therefore turn to a model-based analysis that disentangles the different components underlying free recall.

Model-Based Analyses

We tallied the frequencies of the observed recall events E_1 to E_4 for each participant and study-test trial. The cognitive modeling was based on these frequencies. We used a reparameterized multitrial version of the pair-clustering MPT model (as described above) to quantify possible changes in cluster formation through repeated list presentation. This model provides (a) an estimate of an *initial value* for each parameter in the first study-test trial and (b) estimates of the *change-rate* for each parameter in subsequent trials.

Participant variability. To examine variability in response frequencies (E_1 to E_4) between participants, we performed χ^2 tests (as proposed by Smith & Batchelder, 2008) within each age group and study-test trial, which indicated substantial heterogeneity: smallest $\chi^2(144) = 179.86$, largest $p < .03$ (except in the children's initial two trials in which performance was generally low and data were consequently relatively homogenous). This indicated the necessity of hierarchical modeling to account for the individual variability in recall.

Model comparison and model fit. In a next step, we ran and compared in each age group two hierarchical MPT-model versions that differed in their specification of change (all other things being equal): one model with constant change rate across trials and another model with flexible change rates that were free to vary between trials. The model equations for both model versions are in the Online Supplement. For younger adults, 10-year-olds, and 7-year-olds, the deviance information criterion (DIC) for these two model versions was 2034 vs. 2066, 1763 vs. 1784, and 1460 vs. 1470, respectively. The DIC is a Bayesian alternative to the Akaike information criterion and quantifies the balance between parametric complexity and goodness of fit to determine relative performance of a model. Models with smaller DIC are preferred over models with larger DIC. According to Spiegelhalter, Thomas, Best, and Lunn (2003), differences in DIC between 5 and 10 are "substantial" and differences of more than 10 indicate to "definitely rule out the model with the higher DIC." This suggests that in the current study, improvements in fit by allowing flexible change rates did not justify the additional parametric model complexity in any age group. Therefore, we used a learning model with only one change rate that was constant across trials (for each parameter) in all subsequent analyses. Posterior predictive checks (including statistical fit indices, Klauer, 2010, and visual inspection of observed data and model predictions) indicated a good model fit to the data for 7-year-olds ($P_{T_1} = .34$; $P_{T_2} = .47$), 10-year-olds ($P_{T_1} = .50$; $P_{T_2} = .24$), and adults ($P_{T_1} = .09$; $P_{T_2} = .24$). Details about model fit, model implementation, parameter recovery, and recall data are in the Appendix and Online Supplemental Materials.

Modeling results. Table 2 (lower section) shows the group-level means and standard deviations of the initial parameter values (c_1, r_1, u_1) and of their change rates (β_c, β_r, β_u). Figure 3 shows the corresponding learning curves (changes in c, r, and u across trials) for the three age groups. These learning curves were calculated from the initial parameter values and change rates of the multitrial model, based on the relation $\theta_n = \theta_{n-1} + \beta_\theta(1 - \theta_{n-1})$, for the value of parameter θ at trial n (with $1 < n \leq 4$). In what follows, we focus on differences between age-group means in the initial parameters (first study-test trial) and in the change rates (learning) across study-test trials. Online Supplement 3 includes additional plots of individual parameter estimates. We consider the credibility of group differences and report 95% credibility intervals (highest posterior density) in brackets, based on the MCMC sampling.

Cluster encoding and retrieval. Figure 3, Panel A, shows learning curves for cluster encoding. Parameter c_1 represents the probability of encoding two semantically related words as a cluster in the first trial. The only credible group-mean difference in c_1 emerged between adults and 7-year-olds, $\Delta c_1 = .057$ [.003, .112], indicating that on the initial trial, adults more likely than 7-year-olds encoded item pairs as clusters. Neither adults and 10-year-olds [−.018, .098], nor 10-year-olds and 7-year-olds [−.024, .061] differed credibly in such initial clustering. Importantly, the age groups differed in change of clustering across study-test trials (change-rate parameter β_c): the probability of clustering word pairs increased more strongly across list repetitions in adults than in 10-year-olds, $\Delta\beta_c = .117$ [.070, .163], and in adults than in 7-year-olds $\Delta\beta_c = .150$ [.107,

.194]. The change rate β_c was also higher in 10-year-olds than in 7-year-olds $\Delta\beta_c = .033$ [.009, .058]. Notably, and in contrast with other age groups, 7-year-olds showed no increase in clustering across trials (see Figure 3A). That is, their change rate of β_c was not credibly different from zero [.000, .025].

Figure 3, Panel B, shows changes in parameter r, the probability of retrieving a stored cluster. Parameter r_1 on the initial study-test trial did not differ between age groups (adults vs. 10-year-olds [−.308, .619], adults vs. 7-year-olds [−.064, .956], 10-year-olds vs. 7-year-olds [−.298, .915]). Similarly, no credible age differences emerged in the change of cluster retrieval across trials: the change rate β_r differed neither between adults and 10-year-olds [−.559, .759], nor adults and 7-year-olds [−.610, .739], nor between 10-year-olds and 7-year-olds [−.836, .607]. If only few clusters are encoded in the first place—as was the case particularly in initial trials for the children—then only few clusters can be subsequently retrieved, resulting in large uncertainty in parameter r.

Singleton encoding and retrieval. Parameter u_1 measures the probability of encoding and retrieving a word as a singleton (i.e., without pair clustering). As parameter u measures a conglomerate of singleton encoding and retrieval, it has received less attention in previous research than parameters c and r. Importantly however, a comparison of the pattern of u with that of the other parameters is informative in the present study because such a comparison can reveal why recall performance changes across trials in children: If increases in the singleton parameter u were more pronounced than increases in the clustering parameter c, this would suggest that clustering was not the main route contributing to increases in children's recall performance with repeated list presentations.

Figure 3, Panel C shows the estimates for the singleton parameter u. The age groups differed in estimates of the initial singleton parameter: u_1 was higher for adults than for 10-year-olds, $\Delta u_1 = .059$ [.022, .098], and higher for adults than 7-year-olds, $\Delta u_1 = .092$ [.055, .129]; moreover, u_1 was higher for 10-year-olds than 7-year-olds, $\Delta u_1 = .033$ [.005, .060]. The change rate β_u was higher for adults than for 10-year-olds, $\Delta\beta_u = .098$ [.054, .141], higher for adults than 7-year-olds $\Delta\beta_u = .161$ [.121, .203], and higher for 10-year-olds than 7-year-olds $\Delta\beta_u = .063$ [.039, .086]. Notably, improvements in remembering singletons were significant in all age groups: Even though 7-year-olds showed relatively subtle improvements across trials, their change rate β_u was credibly different from zero, [.008, .035]. Overall, the modeling results suggest that the increases in recall in this age group were mainly due to remembering items as singletons—and not due to categorical clustering.

Correlations. We also explored the correlations (a) among the model parameters and (b) between recall performance (proportion of recalled of words and of word pairs) and the model parameters. Online Supplement 1 and 2 includes details. In adults, parameter c_1 and β_c and parameters u_1 and β_u were positively correlated (rs > .57), respectively, suggesting that people who started performing at a higher level also had higher learning rates across trials in these parameters. We did not find further credible correlations among model parameters in any age group.

Moreover, the correlations between recalled words and the model parameters indicated positive relations: for adults, the correlations between recall and parameter c_1, β_c, u_1, and β_u, respectively, were all credibly positive (rs > .53), whereas for children, this was the case only for the correlations between recall and parameters u_1 and β_u, respectively (rs > .60).

Taken together, these analyses indicate that the model parameters represent largely independent cognitive components that contribute jointly to observed recall. The correlations also suggest that encoding items as clusters is particularly associated with high recall performance in adults, whereas recall of individual items appears to be the major source contributing to younger children's recall accuracy.

Discussion

In this study, we investigated differences between 7-year-old children, 10-year-old children, and young adults in the clustering of categorically related words in free recall. An important question was whether multiple presentations of a list might increase clustering and help children to improve their performance. To measure cluster formation at encoding and to assess how clustering supported free recall performance, we used a Bayesian hierarchical MPT model that allowed us to disentangle different cognitive components underlying free recall. To our knowledge, this is the first study to use a cognitive modeling approach to illuminate children's clustering in free recall.

The main findings can be summarized as follows. First, on a behavioral level, the age groups expectedly differed from each other in recall of total number of words and of word pairs, with adults showing the highest and younger children showing the lowest performance. These behavioral differences were accounted for by corresponding age differences in the encoding of categorically related words as clusters (model parameter c) and in remembering items as singletons (parameter u). Interestingly, the modeling indicated that the age groups differed from the outset in memory for singletons, whereas age differences in initial cluster encoding were less clear-cut (only adults and 7-year-olds differed credibly in parameter c_1). This suggests that clustering-strategy development is not the only factor contributing to age differences in recall and that even young adults needed more than one trial to utilize categorical relations between items. Notably, developmental differences in cluster encoding transpired mostly in the change rate across trials (parameter β_c). This highlights that age differences in cognitive processes can be more pronounced in learning than in the initial baseline level (cf. Bröder et al., 2008). Estimates of the conditional probability of retrieving clusters from memory, given their successful encoding, were generally high (parameter r). This appears to suggest that once items were successfully encoded and stored together, there was a high likelihood of retrieving the cluster again, and there were no credible age differences in cluster retrieval (see Howe et al., 1985; Howe & O'Sullivan, 1997, for similar findings in other experimental memory paradigms). However, these quantitative invariances in retrieval should be interpreted with caution in the present study: Because few clusters were encoded on initial trials in the first place (particularly by children), the uncertainty in this parameter was relatively large, and we therefore refrain from stronger conclusions about cluster retrieval. On the whole, and in line with previous research, these findings indicate clear developmental differences in the semantic organization of episodic memory which could be localized in the ability to cluster categorically related information at encoding.

Second, all age groups showed significant increases across study-test trials in word recall. However, increases were generally most pronounced in the adults and least pronounced in the younger children. We found Age × Trial interactions in recall performance and in the model parameters, following a pattern that we described as *Scenario C* in the introduction. That is, age differences increased, rather than decreased, as learning progressed through repeated study-test cycles (cf. Cole et al., 1971). In 7-year-olds, initial cluster encoding was close to zero and, in contrast to the other age groups, this group showed no increase in cluster encoding across trials (parameter β_c). Hence, repeated study opportunities did not help 7-year-olds to encode categorically related items as clusters, but all age groups showed increases in remembering singletons (parameter β_u). The larger age differences in the change rate of cluster encoding (β_c) than of singleton memory (β_u) suggest that acquiring proficiency to form relations among items is more difficult for younger children than improving memory for individual items (Sloutsky & Fisher, 2004; Schneider, 2014).

Third, even though learning rates differed substantially between groups, the *shape* of learning in free-recall clustering followed a common regularity in all age groups, indicating structural developmental invariance. That is, in each age group, a model assuming a constant change rate across trials (implying geometric growth) fit the data very well and outperformed a more flexible model version (in which change was unconstrained across trials). Models with a constant change rate have been successfully applied in many other domains of learning (for overviews, see Brainerd, 1983; Bush & Mosteller, 1955). Here, we could show that such models also adequately describe the learning of clustering in free recall, pointing to a general learning principle across age groups.

What are implications of the current findings from a developmental perspective? Recall improved to some extent through repeated learning opportunities in all age groups. In younger children, however, the effect of this manipulation on clustering-strategy use was negligible: Whereas adults' categorical organization of related words increased substantially over trials, 7-year-old children showed no change in the use of the category structure of the word list. Hence, their moderate improvements in recall stemmed from encoding and retrieving singleton items (as measured by parameter *u*) and thus may have been achieved mainly through memorizing items in a non-strategic, rote manner. Moreover, the initial probability of cluster encoding in the younger age groups was lower than the initial probability of remembering singletons, pointing to the latter as a main route contributing to children's recall. These results dovetail with past research that showed that even younger children may have good memory for individual items (Sloutsky & Fisher, 2004) and that the basic ability for long-term recall develops from an early age (e.g., Bauer et al., 2003), whereas the formation of semantic relations between items undergoes protracted development (Bjorklund, 2011; Schneider, 2014). Taken together, our findings clearly indicate that age differences in categorical clustering cannot be reduced (or even eliminated) by merely providing more experience with a recall task. Instead, more intensive or different interventions (e.g., training of mnemonic strategy use) may be necessary. Indeed, research has repeatedly found that even younger children, who show little evidence of spontaneous organization, can be trained to cluster information under certain instructional conditions and to thus increase their memory performance (Bjorklund, 2011; Moely et al., 1969; Rao & Moely, 1989). In other words, younger children appear capable of organizing information for recall, but they generally fail to do so spontaneously. Our findings extend this literature by showing that study-test repetitions alone are not sufficient to overcome such a production deficiency in younger children.

The current findings also provide important information for other lines of research. For example, the development of meaning connection (i.e., improvements in the ability to make meaningful connections between experiences) may contribute to susceptibility to false memory. Prominent theories of memory development such as fuzzy-trace theory (e.g., Reyna & Brainerd, 1995) predict age-related increases in false remembering across a variety of tasks (intrusions in recall or false alarms in recognition) based on the developing ability to form semantic relations across exemplars and the tendency to rely more on qualitative, meaning-based representations from memory (so-called gist traces). Several studies have reported the seemingly counterintuitive finding that false-memory phenomena are relatively amplified with increasing age and that these trends are encoding- or storage-driven rather than being retrieval effects (for overviews, see Brainerd, Reyna, & Ceci, 2008, and Sloutsky & Fisher, 2004). However, Brainerd et al. (p. 349) emphasized the importance of defining and measuring meaning connection independently of false-memory phenomena and to verify developmental trends in meaning connection independently. The current study provides such an investigation with a paradigm that does not involve induction of false memories. Consistent with research using false-memory paradigms, our findings suggest that the formation of meaningful connections at encoding and their contributions to memory performance increase substantially over the elementary-school years and into adulthood.

There are also limitations that could not be addressed in the current study. First, the present findings are based on cross-sectional data and await replication with longitudinal designs. Second, we did not examine in greater depth the role of core cognitive abilities (e.g., fluid abilities, working memory), which may also partly explain individual and developmental differences in clustering and free recall. In the current study, we did not find systematic relations between digit span (backward and forward) and the MPT-model parameters (the credibility intervals of all of these correlations included zero). However, future research could examine these issues in greater detail through more extensive cognitive testing. Finally, we took several steps (pilot testing, independent teacher ratings, auditory presentation) to ensure that even first graders would be familiar with all presented words and understand all instructions. Nonetheless, we cannot exclude that older children were generally more proficient than younger children in dealing with verbal material; it could thus be interesting to examine clustering with other types of materials in the future (e.g., pictures or toy objects).

Conclusion

We conducted a cognitive-modeling analysis of category clustering in free recall by school-age children and adults. In all age groups, repeated study-test trials improved recall and learning followed a common pattern, suggesting developmental invariance in the way category clustering changed across trials. In contrast to older age groups, moderate increases in recall of 7-year-olds were exclusively based on remembering individual items—and not on encoding them as clusters. Our research thus highlights the potential and regularity of verbal learning across childhood, but also suggests that repeated learning opportunities are insufficient to induce clustering strategies in younger children.

References

Ackerman, B. P. (1981). Encoding specificity in the recall of pictures and words in children and adults. *Journal of Experimental Child Psychology, 31,* 193–211.

Arnold, N. R., Bayen, U. J., & Böhm, M. F. (2015). Is prospective memory related to depression and anxiety? A hierarchical MPT modelling approach. *Memory, 23,* 1215–1228.

Batchelder, W. H., & Riefer, D. M. (1980). Separation of storage and retrieval factors in free recall of clusterable pairs. *Psychological Review, 87*, 375–397.

Batchelder, W. H., & Riefer, D. M. (1986). The statistical analysis of a model for storage and retrieval processes in human memory. *British Journal of Mathematical and Statistical Psychology, 39*, 129–149.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6,* 57–86.

Bates, E., D'Amico, S., Jacobsen, T., Székely, A., Andonova, E., Devescovi, A., …Tzeng, O. (2003). Timed picture naming in seven languages. *Psychonomic Bulletin & Review, 10*, 344–380.

Bauer, P. J., Wiebe, S. A., Carver, L. J., Waters, J. M., & Nelson, C. A. (2003). Developments in long-term explicit memory late in the first year of life: Behavioral and electrophysiological indices. *Psychological Science, 14,* 629–635.

Bender, R. H., Wallsten, T. S., & Ornstein, P. A. (1996). Age differences in encoding and retrieving details of a pediatric examination. *Psychonomic Bulletin & Review, 3,* 188–198.

Bäuml, K.-H. (1991). Experimental analysis of storage and retrieval processes involved in retroactive inhibition: The effect of presentation mode. *Acta Psychologica, 77,* 103–119.

Bernstein, D. M., Erdfelder, E., Meltzoff, A. N., Peria, W., & Loftus, G. R. (2011). Hindsight bias from 3 to 95 years of age. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 378–391. doi: 10.1037/a0021971

Bjorklund, D. F. (2011). *Children's thinking: Cognitive development and individual differences.* Belmont, CA: Wadsworth.

Bjorklund, D. F., & Jacobs III, J. W. (1985). Associative and categorical processes in children's memory: The role of automaticity in the development of organization in free recall. *Journal of Experimental Child Psychology, 39*, 599–617.

Bousfield, A. K., & Bousfield, W. A. (1966). Measurement of clustering and of sequential constancies in repeated free recall. *Psychological Reports, 19,* 935–942.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive Psychology, 1,* 18–46.

Brainerd, C. J. (1983). Structural invariance in the developmental analysis of learning. In J. Bisanz, G. L. Bisanz, & R. Kail (Eds.), *Learning in children* (pp. 1-36). New York: Springer.

Brainerd, C. J. (1985). Model-based approaches to storage and retrieval development. In C. J. Brainerd & M. Pressley (Eds.), *Basic processes in memory development: Progress in cognitive development research* (pp. 143–208). New York: Springer.

Brainerd, C. J., Howe, M. L., Kingma, J., & Brainerd, S. H. (1984). On the measurement of storage and retrieval contributions to memory development. *Journal of Experimental Child Psychology, 37,* 478–499.

Brainerd, C. J., Reyna, V. F., & Ceci, S. J. (2008). Developmental reversals in false memory: A review of data and theory. *Psychological Bulletin, 134,* 343–382.

Bröder, A., Herwig, A., Teipel, S., & Fast, K. (2008). Different storage and retrieval deficits in normal aging and mild cognitive impairment: A multinomial modeling analysis. *Psychology and Aging, 23*, 353–365.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models of learning.* New York: Wiley.

Chechile, R. A., Richman, C. L., Topinka, C., & Ehrensbeck, K. (1981). A developmental study of the storage and retrieval of information. *Child Development, 52,* 251–259.

Cole, M., Frankel, F., & Sharp, D. (1971). Development of free recall learning in children. *Developmental Psychology, 4,* 109–123.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Journal of Psychology, 217,* 108–124.

Erdfelder, E., & Bayen, U. J. (1991). Episodisches Gedächtnis im Alter: Methodologische und empirische Argumente für einen Zugang über mathematische Modelle [Episodic memory in old age: Methodological and empirical arguments for a mathematical modeling approach]. In D. Frey (Ed.), *Bericht über den 37. Kongreß der Deutschen Gesellschaft für Psychologie in Kiel 1990* [Proceedings of the 37th Conference of the German Society for Psychology in Kiel 1990] (Vol. 2, pp. 172-180). Göttingen, Germany: Hogrefe.

Fandakova, Y., Lindenberger, U., & Shing, Y. L. (2015). Episodic memory across the lifespan: General trajectories and modifiers. In D. R. Addis, M. Barense, & A. Duarte (Eds.), *Wiley handbook on the cognitive neuroscience of memory* (pp. 309–325). Hoboken: Wiley-Blackwell.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191.

Filevich, E., Horn, S. S., & Kühn, S. (2019). Within-person adaptivity in frugal judgments from memory. *Psychological Research, 83,* 613–630.

Ghetti, S., & Lee, J. (2011). Children's episodic memory. *Wiley Interdisciplinary Reviews: Cognitive Science, 2,* 365–373.

Glidden, L. M. (1977). Developmental effects in free recall learning. *Child Development, 48,* 9–12.

Halford, G. S., Wilson, W. H., & Phillips, S. (1998). Processing capacity defined by relational complexity: Implications for comparative, developmental, and cognitive psychology. *Behavioral and Brain Sciences, 21,* 803–865.

Hasselhorn, M. (1990). The emergence of strategic knowledge activation in categorical clustering during retrieval. *Journal of Experimental Child Psychology, 50*, 59–80.

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An R package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods, 50,* 264–284.

Horn, S. S., Pachur, T., & Mata, R. (2015). How does aging affect recognition-based inference? A hierarchical Bayesian modeling approach. *Acta Psychologica, 154,* 77–85.

Horn, S. S., Ruggeri, A., & Pachur, T. (2016). The development of adaptive decision making: Recognition-based inference in children and adolescents. *Developmental Psychology, 52,* 1470–1485.

Howe, M.L., Brainerd, C.J., & Kingma, J. (1985). Development of organization in recall: A stages-of learning analysis. *Journal of Experimental Child Psychology, 39,* 230–251.

Howe, M. L., & O'Sullivan, J. T. (1997). What children's memories tell us about recalling our childhood: A review of storage and retrieval processes in the development of long-term retention. *Developmental Review, 17,* 148–204.

Jablonski, E. M. (1974). Free recall in children. *Psychological Bulletin, 81,* 522–539.

Klauer, K. C. (2010). Hierarchical multinomial processing tree models: A latent-trait approach. *Psychometrika, 75,* 70–98.

Knapp, B. R., & Batchelder, W. H. (2004). Representing parametric order constraints in multi-trial applications of multinomial processing tree models. *Journal of Mathematical Psychology, 48,* 215–229.

Lee, M. D. (2018). Bayesian methods in cognitive modeling. In J. Wixted & E.-J. Wagenmakers (Eds.), *The Stevens Handbook of Experimental Psychology and Cognitive Neuroscience, Volume 5: Methodology*. Hoboken, NJ: John Wiley & Sons.

Lee, M. D., & Wagenmakers, E. J. (2014). *Bayesian cognitive modeling: A practical course.* Cambridge University Press.

Lehmann, M., & Hasselhorn, M. (2010). The dynamics of free recall and their relation to rehearsal between 8 and 10 years of age. *Child Development, 81,* 1006–1020.

Mandler, G. (1967). Organization and memory. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation.* Vol. 1. New York: Academic Press.

Mannhaupt, H.-R. (1983). Produktionsnormen für verbale Reaktionen zu 40 geläufigen Kategorien. *Sprache und Kognition, 4,* 264–278.

Matzke, D., Dolan, C. V., Batchelder, W. H., & Wagenmakers, E. J. (2015). Bayesian estimation of multinomial processing tree models with heterogeneity in participants and items. *Psychometrika, 80,* 205–235.

Michalkiewicz, M., & Erdfelder, E. (2016). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition, 44,* 454–468.

Mizrak, E., Singmann, H., & Öztekin, I. (2017). Forgetting emotional material in working memory. *Social Cognitive and Affective Neuroscience, 13,* 331–340.

Moely, B. E., Olson, F. A., Halwes, T. G., & Flavell, J. H. (1969). Production deficiency in young children's clustered recall. *Developmental Psychology, 1,* 26–34.

Moely, B. E., & Shapiro, S. I. (1971). Free recall and clustering at four age levels: Effects of learning to learn and presentation method. *Developmental Psychology, 4,* 490.

Ornstein, P. A., Naus, M. J., & Liberty, C. (1975). Rehearsal and organizational processes in children's memory. *Child Development, 46,* 818–830.

Paris, S. G. (1978). Memory organization during children's repeated recall. *Developmental Psychology, 14,* 99–106.

Petermann, F. & Petermann, U. (2007). *Hamburg-Wechsler-Intelligenztest für Kinder IV (HAWIK-IV).* [Wechsler Intelligence Scale for Children, Version IV, German Adaptation]. Göttingen, Germany: Hogrefe.

Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing* (Vol. 124, No. 125.10).

Pohl, R. F., Bayen, U. J., Arnold, N., Auer, T.-S., & Martin, C. (2018). Age differences in processes underlying hindsight bias: A life-span study. *Journal of Cognition and Development, 19,* 278–300.

Pohl, R. F., Bayen, U. J., & Martin, C. (2010). A multi-process account of hindsight bias in children. *Developmental Psychology, 46,* 1268–1282.

Rao, N., & Moely, B. E. (1989). Producing memory strategy maintenance and generalization by explicit or implicit training of memory knowledge. *Journal of Experimental Child Psychology, 48,* 335–352.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-Trace theory: An interim synthesis. *Learning and Individual Differences, 7,* 1–75.

Riefer, D. M. (1982). The advantages of mathematical modeling over traditional methods in the analysis of category clustering. *Journal of Mathematical Psychology, 26,* 97–123.

Riefer, D. M., & Batchelder, W. H. (1991). Age differences in storage and retrieval: A multinomial modeling analysis. *Bulletin of the Psychonomic Society, 29*, 415–418.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment, 14*, 184–201.

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2016). Hierarchical Bayesian models. In W. H. Batchelder, H. Colonius, E. Dzhafarov, and J. I. Myung (Eds.), *New Handbook of Mathematical Psychology: Volume 1, Foundations and Methodology*. Cambridge University Press.

Schacter, D. L., & Tulving, E. (1994). *Memory systems 1994* (1st ed.). Cambridge, MA: MIT Press.

Schaper, M. L., Kuhlmann, B. G., & Bayen, U. J. (2018). Metamemory expectancy illusion and schema-consistent guessing in source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Online first publication. doi: 10.1037/xlm0000602

Schneider, W. (2014). *Memory development from early childhood through emerging adulthood.* Heidelberg, Germany: Springer.

Sloutsky, V. M., & Fisher, A. V. (2004). When development and learning decrease memory: Evidence against category-based induction in children. *Psychological Science, 15,* 553–558.

Smith, J. B., & Batchelder, W. H. (2008). Assessing individual differences in categorical data. *Psychonomic Bulletin & Review, 15,* 713–731.

Smith, R. E., Bayen, U. J., & Martin, C. (2010). The cognitive processes underlying event-based prospective memory in school age children and young adults: A formal model-based study. *Developmental Psychology, 46,* 230–244.

Spiegelhalter, D., Thomas, A., Best, N., & Lunn, D. (2003). *WinBUGS user manual.* Cambridge, UK: MRC Biostatistics Unit.

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin & Review,* 25, 143–154.

von Aster, M., Neubauer, A., & Horn, R. (2006). *WIE III. Wechsler Intelligenztest für Erwachsene.* [Wechsler Adult Intelligence Scale, German Adaptation, Version III].

Yim, H., Dennis, S. J., & Sloutsky, V. M. (2013). The development of episodic memory: Items, contexts, and relations. *Psychological Science, 24,* 2163–2172.

Table 1

Participant Characteristics and Cognitive Test Scores

	7-year-olds $n = 49$ 27*f*; 22*m*		10-year-olds $n = 49$ 27*f*; 22*m*		Adults $n = 52$ 34*f*; 18*m*	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age (years)	6.90	0.34	10.11	0.49	21.42	2.86
Similarities	10.96	3.56	10.73	2.02	9.62	1.75
Digit span	10.20	2.65	10.76	2.28	10.60	2.39
Digit-symbol coding	11.16	2.70	11.63	2.48	11.48	2.84

Notes. f = female; m = male. For children, the cognitive tests were taken from the German version of the Wechsler Intelligence Scale for Children (Petermann & Petermann, 2007); for adults, the corresponding tests were taken from the German version of the Wechsler Adult Intelligence Scale (von Aster, Neubauer, & Horn, 2006). Age-scaled norm scores from Wechsler subtests have $M = 10$ and $SD = 3$. For the digit span subtest, the raw scores from the forward-span and backward-span tasks were summed and then transformed into a norm score. Analyses of cognitive test scores indicated that participants performed at least as well as normatively expected for their respective age group [smallest $t(51) = -1.58$; $p > .11$; for similarities scale, adults].

Table 2

Proportions of Recalled Words, Proportions of Recalled Word Pairs, and Multinomial-Model Parameters for each Age Group

		M (SD)		
		7-year-olds	10-year-olds	Adults
Word recall				
	Trial			
	#1	.13 (.05)	.19 (.07)	.29 (.13)
	#2	.19 (.08)	.28 (.08)	.50 (.17)
	#3	.23 (.10)	.38 (.12)	.65 (.16)
	#4	.23 (.12)	.45 (.14)	.77 (.15)
	overall	.19 (.07)	.32 (.09)	.55 (.13)
Word-pair recall				
	Trial			
	#1	.02 (.04)	.06 (.06)	.13 (.12)
	#2	.06 (.06)	.11 (.08)	.33 (.21)
	#3	.07 (.07)	.19 (.11)	.50 (.19)
	#4	.09 (.09)	.28 (.14)	.66 (.20)
	overall	.06 (.05)	.16 (.07)	.41 (.16)
Model parameters				
cluster encoding *c*				
	initial value c_1	.034 (.013)	.054 (.024)	.109 (.084)
	change rate β_c	.015 (.005)	.050 (.021)	.190 (.125)
cluster retrieval *r*				
	initial value r_1	.454 (.422)	.612 (.320)	.835 (.119)
	change rate β_r	.499 (.442)	.445 (.371)	.533 (.214)
singleton recall *u*				
	initial value u_1	.131 (.034)	.163 (.030)	.228 (.075)
	change rate β_u	.033 (.032)	.091 (.038)	.199 (.100)

Note. Word recall $Pr(C_1)$ and word-pair recall $Pr(C_2)$ can be calculated from the proportion of responses in the model categories as $Pr(C_1) = Pr(E_1) + Pr(E_2) + Pr(E_3)/2$ and $Pr(C_2) = Pr(E_1) + Pr(E_2)$. The model-parameter estimates are the group-level means and group-level standard deviations (in parentheses) of the hierarchical latent-trait model. The initial value (θ_1) is the parameter estimate of the first study-test trial; the corresponding change-rate β_θ is the proportional reduction of error from one trial to the next. Group-level means and standard deviations of the MPT model parameters are reported on the probability scale (ranging from 0 to 1). For this purpose, we applied the inverse probit transformation on all latent-trait-model MCMC estimates, using the probitInverse() function in the TreeBUGS package for *R* (Heck, Arnold, & Arnold, 2018).

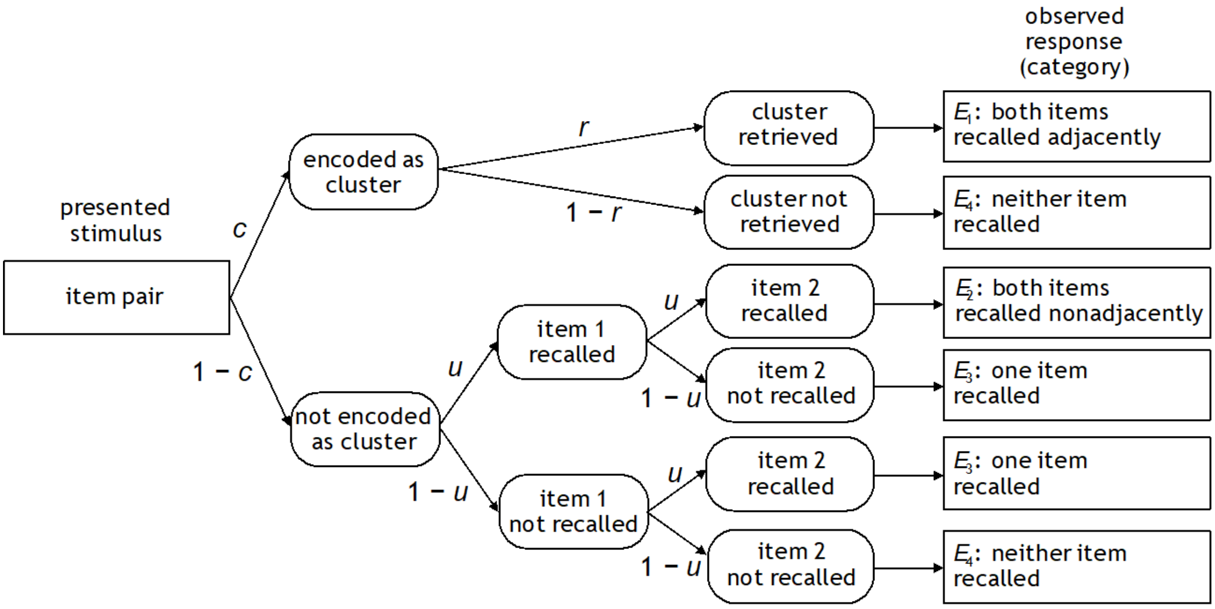


Figure 1. The pair-clustering multinomial model of free recall (Batchelder & Riefer, 1980, 1986) for item lists containing clusterable pairs. c ≡ probability of encoding and storing related items as a cluster; r ≡ probability of retrieving a cluster at test; u ≡ probability of recalling (encoding and retrieving) an item as a singleton. Possible free-recall events (categories E_1 to E_4) are represented as rectangles on the right side in the figure.

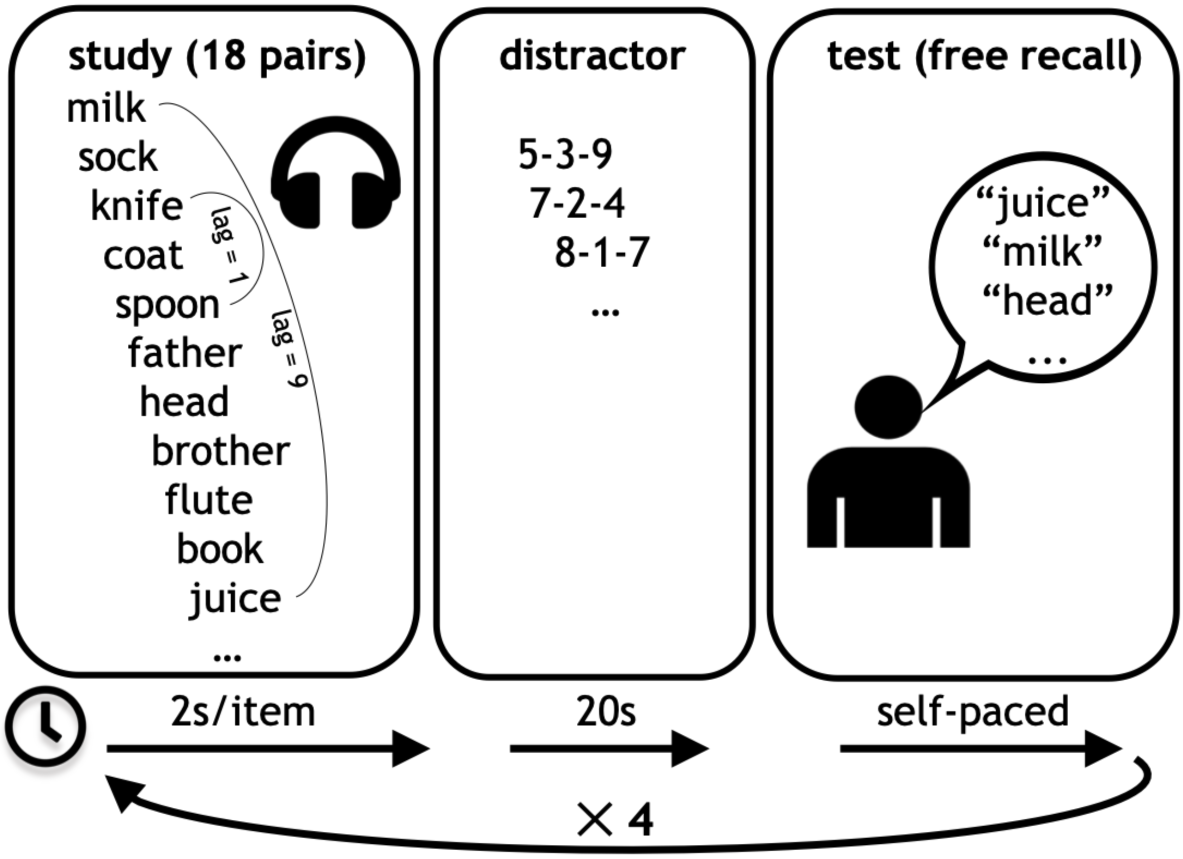
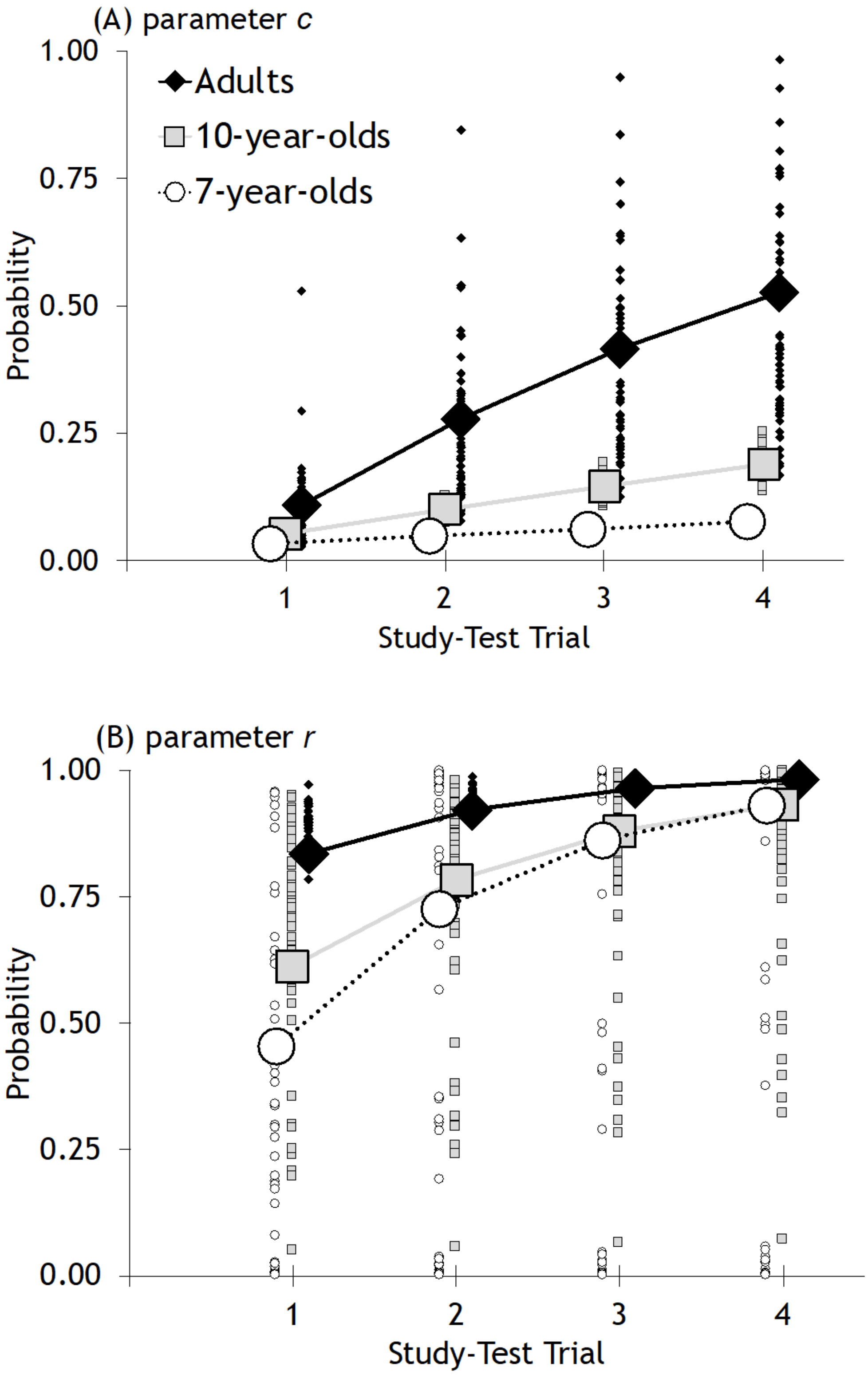


Figure 2. Illustration of the study-test procedure. Each participant completed four study-test trials in which the same eighteen pairs of categorically related words were presented during each study phase, followed by a distractor task and a free recall phase.


(A) parameter *c*
1.00
0.75
0.50
0.25
0.00
Probability
Adults
10-year-olds
7-year-olds
1
2
3
4
Study-Test Trial
(B) parameter *r*
1.00
0.75
0.50
0.25
0.00
Probability
1
2
3
4
Study-Test Trial

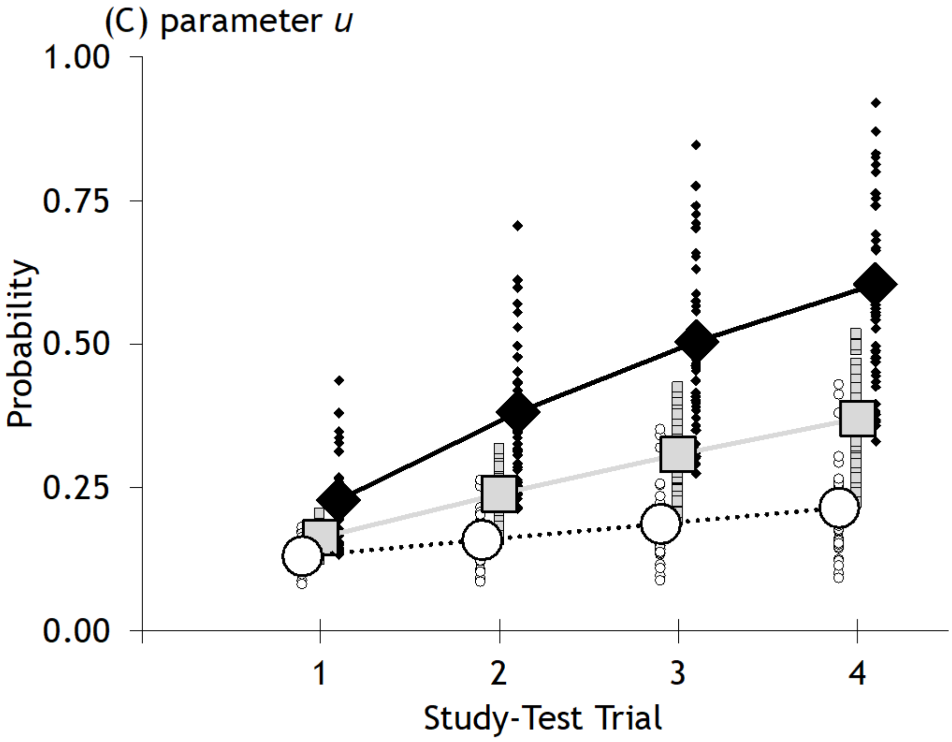


Figure 3. Learning curves as a function of age group and study-test trial, calculated from the parameter estimates of the multitrial multinomial model with order constraints on the parameters and constant change rates across trials (Bush-Mosteller learning model). In addition to the group-level curves (larger symbols), the individual estimates are also shown (smaller symbols). Panel A: probability of encoding and storing a word pair as a cluster (parameter *c*). Panel B: probability of retrieving a cluster at test (parameter *r*). Panel C: probability of recalling a word as a singleton (parameter *u*). Within each study-test trial, an offset between groups is included to visually distinguish the values for the different groups.

Appendix A
Hierarchical Bayesian Model Implementation

Latent-Trait Model Approach

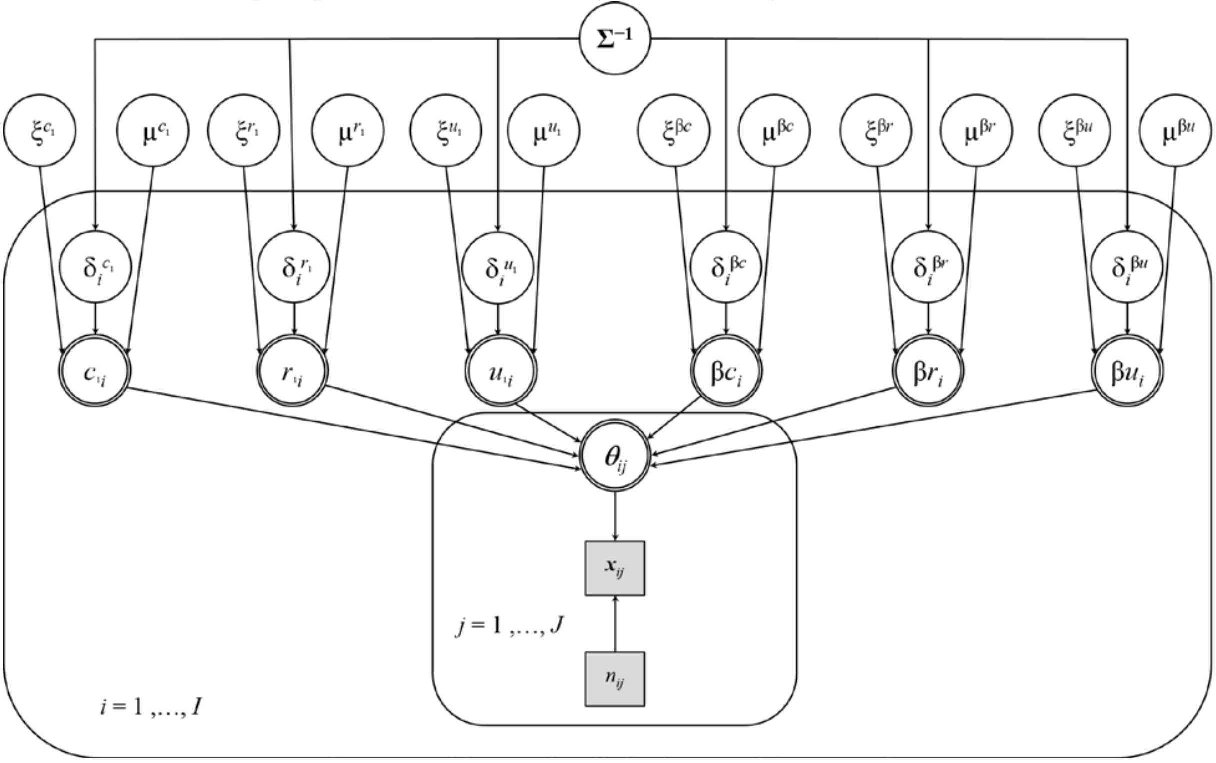
We used a hierarchical (multi-level) MPT model that can account for individual differences in the cognitive components underlying observed behavior. An advantage of this so-called *latent-trait model* approach (Klauer, 2010) is that both group-level and individual-level parameters are obtained in a principled way, which makes it also possible to jointly estimate the correlations between parameters and with external covariates (e.g., scores from other cognitive tests). We closely followed the implementations as described in Klauer (2010) and Matzke et al. (2015).

Bayesian Parameter Estimation

The estimation of the model parameters relied on Bayesian inference (for overviews, see Lee, 2018; Lee & Wagenmakers, 2014; Rouder, Morey, & Pratte, 2016) which has been applied in many areas of cognitive modeling (e.g., Arnold, Bayen, & Böhm, 2014; Filevich, Horn, & Kühn, 2019; Heck, Arnold, & Arnold, 2017; Horn, Pachur, & Mata, 2015; Kellen, Pachur, & Hertwig, 2016; Michalkiewicz & Erdfelder, 2016; Mizrak, Singmann, & Öztekin, 2017; Schaper, Kuhlmann, & Bayen, 2018). We used the Markov chain Monte Carlo (MCMC) methodology for posterior sampling to determine the most credible value ranges of the model parameters in the posterior distributions given the data. For MCMC sampling with JAGS (Plummer, 2003), we ran three chains of 100,000 iterations each with a thinning rate of 10 and discarded the first 50% of iterations as burn-in. For all model estimates, we report the medians of the MCMC samples.

Graphical Model

The graph structure in Figure A1 illustrates the hierarchical latent-trait version of the re-parameterized pair-clustering model with order constraints (Knapp & Batchelder, 2004) and constant change rates (i.e., we estimated one initial value and one change rate for each parameter of the model; the model equations for this specific implementation are in the Online Supplement). A model for one experimental group is shown; we implemented three such models to estimate and compare parameters between 7-year-olds, 10-year-olds, and adults.



The Figure shows dependencies (probabilistic and deterministic relations) between latent model parameters and the data. Following conventional notation, observed variables are symbolized by shaded nodes, latent variables by unshaded nodes, continuous variables by circular nodes, and discrete variables by square nodes (see Lee & Wagenmakers, 2014). The plates indicate replications over the $J = 4$ different model trees (the four study-test trials in the free-recall paradigm) and over I individuals. For each individual i, the response data $\boldsymbol{x}_{\text{ij}}$ (a vector with a participant's category counts in a tree) follow a multinomial distribution with category probabilities $\boldsymbol{\Theta}_{\text{ij}}$ and number of observations n_{ij}, as defined in the pair-clustering model with order constraints (see Figure 1 for the tree structure).

The individual-level parameters $\boldsymbol{\pi}_i = (c_{1i}, r_{1i}, u_{1i}, \beta c_i, \beta r_i, \beta u_i)$ are modeled in probit-transformed space as $\pi_{\text{i}} \leftarrow \boldsymbol{\Phi}\ (\mu^{\pi} + \delta_i^{\pi} \cdot \xi^{\pi})$ and hence represent linear combinations of a group-level mean μ^{π}, an individual displacement parameter δ_i^{π}, and a multiplicative scaling parameter ξ^{π} (which is redundant, but serves to improve the convergence of the MCMC sampling process). For the Bayesian implementation in the current study, the following prior distributions were specified for the parameters: $\mu^{\pi} \sim N(0,1)$ and $\xi^{\pi} \sim U(0,10)$, where N is the Gaussian and U the uniform distribution. Moreover, individual displacement parameters δ_i^{π} are drawn from a zero-centered multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}^{-1} \sim W(\mathbf{I}_K,$

K +1), where W is the Wishart distribution with K + 1 degrees of freedom and with identity matrix $\mathbf{I}_K$ that has K rows and columns, respectively (the modeling in the current study involved K = 6 individual parameters). For parameter estimation, parameter comparisons, and evaluation of model fit, we used the TreeBugs package for *R* by Heck, Arnold, and Arnold (2018).

Appendix B
Model Fit and Convergence

Fit Indices

To evaluate fit of the observed data to the predictions of the latent-trait model, we examined the fit indices T_1 and T_2, based on posterior model checks (see Klauer, 2010, for details). Index T_1 quantifies the adequacy of a model in accounting for the mean observed response frequencies across model categories, whereas index T_2 quantifies the adequacy of a model in accounting for the variability (variances and covariances) among the observed response frequencies. Posterior predictive *p* values for both T_1 and T_2 indicated a good fit of the model for 7-year-olds (P_{T_1} = .34; P_{T_2} = .47), 10-year-olds (P_{T_1} = .50; P_{T_2} = .24), and adults (P_{T_1} = .09; P_{T_2} = .24). A person-wise examination of individual T_1 indices also indicated good fit in each age group. Overall, these analyses suggested that the version of the pair clustering MPT model we applied (which assumes constant change rate; Bush & Mosteller, 1955) accounted well for the multitrial recall data. This was also supported by a graphical inspection of the data and the posterior predictive distributions, which are plotted below.

Observed Data and Posterior-Predictive Distributions

Figures B1 to B3 show plots of the model predictions (box plots show samples from the model posterior distribution, based on n = 2500 samples) and observed responses (means of individual response frequencies; red triangles) as a function of category in the MPT model. The entries E_{ij} on the *x*-axis refer to the response category $i = 1,\ldots,4$ in the pair-clustering model (see Figure 1 in the article) in study-test trial *j* (with $j = 1,\ldots,4$). Graphs were created with the TreeBugs package (Heck et al., 2018).

Figure B1. Response Frequencies of 7-Year-Old Children.

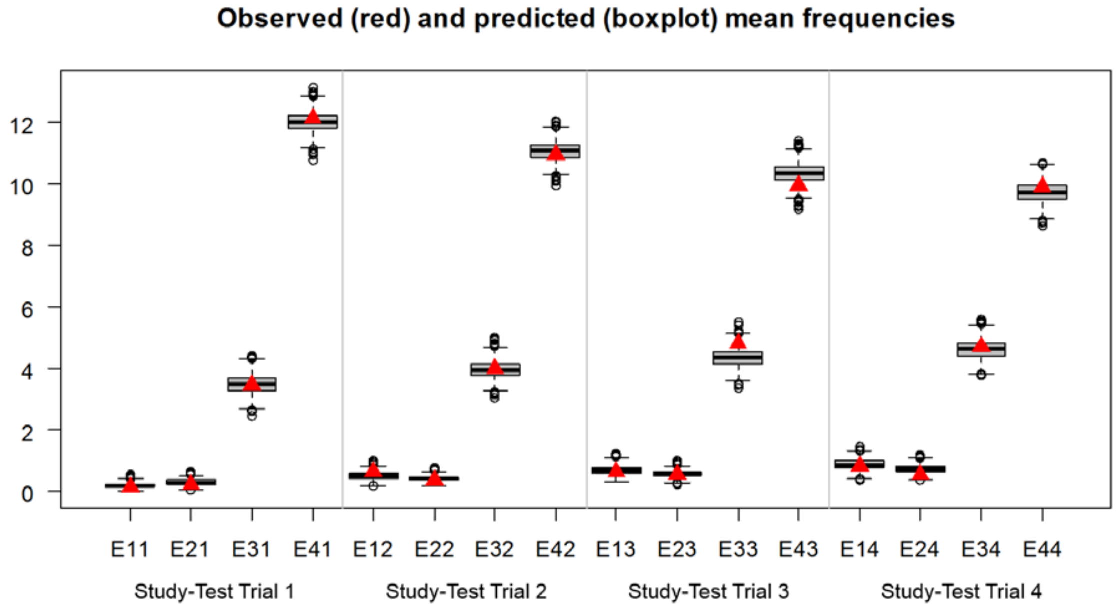


Figure B2. Response Frequencies of 10-Year-Old Children.

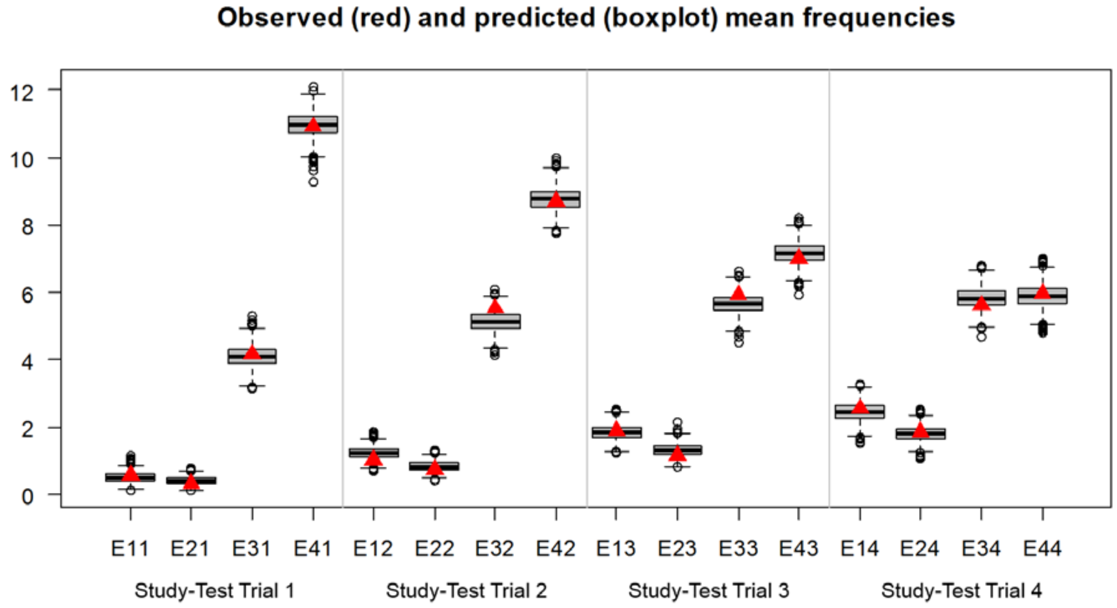
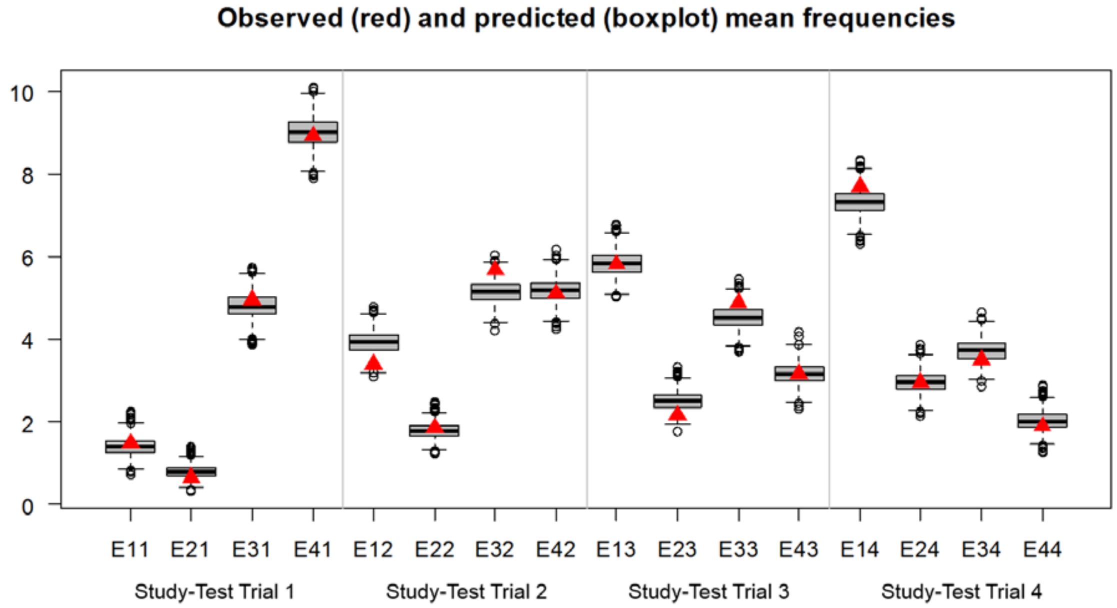


Figure B3. Response Frequencies of Young Adults.



MCMC Chain Convergence

For parameter estimation, we ran three MCMC chains of 100,000 iterations each (with a thinning rate of 10; the first 50% of iterations were discarded as burn-in). Chain convergence was satisfactory for all estimated group-level model parameters ($\hat{R}$ < 1.02) and individual-level parameters ($\hat{R}$ < 1.1). A comprehensive listing of all individual and group-level parameters—including convergence statistics—can be found in the text file *results modelparameters.txt* at *https://tinyurl.com/devpsychrecall*

Appendix C: Different Hypothetical Learning Scenarios

Figures C1 to C3 show three different hypothetical trajectories that illustrate how learning to cluster may progress across study-test trials. The curves were calculated from the Bush-Mosteller (1955) learning model, assuming the same initial values (first trial) for θ_1 across scenarios of .05, 0.125, and 0.25 for 7-year-olds, 10-year-olds, and adults, respectively. The scenarios differ in the rate of learning, as quantified by the proportional change (β_θ) from one trial to the next.

Figure C1. Scenario A: The rate of learning β_θ is the same for all age groups.

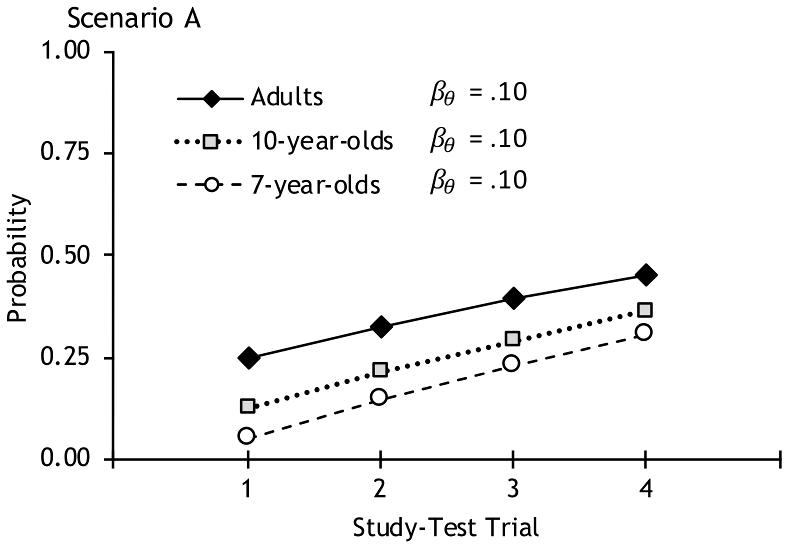


Figure C2. Scenario B: The rate of learning is highest for younger children and lowest for adults.

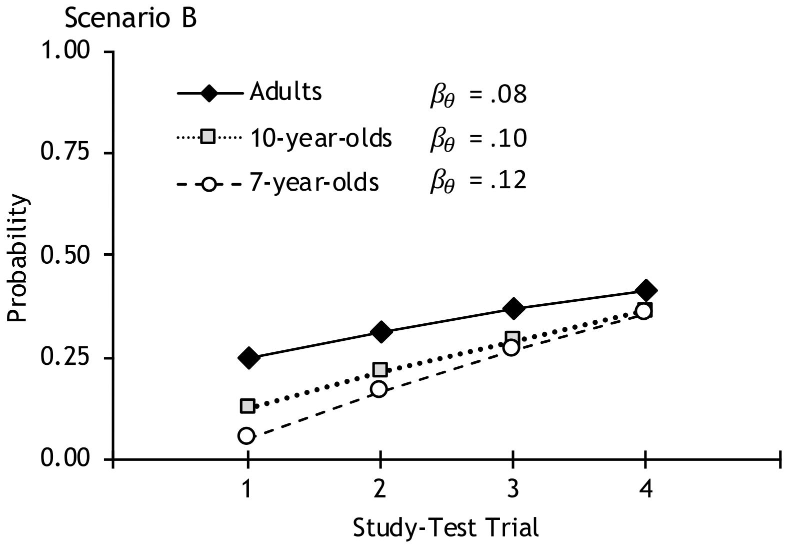
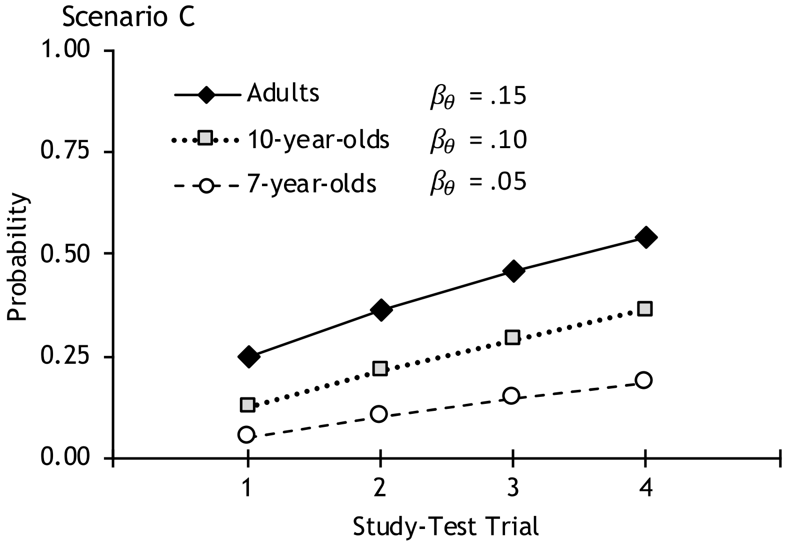


Figure C3. Scenario C: The rate of learning is highest for adults and lowest for younger children.


Scenario C
Probability
Adults β_θ = .15
10-year-olds β_θ = .10
7-year-olds β_θ = .05
Study-Test Trial