Citation for final published version:

Overkott, Clara, Souza, Alessandra S. and Morey, Candice C. ORCID: https://orcid.org/0000-0002-7644-5239 2022. The developing impact of verbal labels on visual memories in children. Journal of Experimental Psychology: General file

Author-final version accepted for publication at the *Journal of Experimental Psychology: General*

The developing impact of verbal labels on visual memories in children

Clara Overkott[1, 2], Alessandra S. Souza[1,3] & Candice C. Morey[2]

[1]University of Zurich, [2]Cardiff University, [3]University of Porto

Author Note

Clara Overkott, Department of Psychology, University of Zurich, Switzerland and School of Psychology, Cardiff University, United Kingdom. Alessandra S. Souza, Faculty of Psychology and Education Sciences, University of Porto, Portugal and Department of Psychology, University of Zurich, Switzerland. Candice C. Morey, School of Psychology, Cardiff University, United Kingdom. This research was support by two grants from the Swiss National Science Foundation (P1ZHP1_187671 and 100019_169302) awarded to C. Overkott and A. S. Souza, respectively, and supported by the Cardiff University Centre for Human Developmental Sciences. We are grateful to the staff of Holton Primary School for their support and to all the families who assisted with the research. The experiment reported in this paper was pre-registered in the Open Science Framework (OSF). Pre-registrations, materials, data, and analysis scripts can be found on: https://osf.io/3wesd/ (Overkott, Morey, & Souza, 2022).

C. Overkott proposed the study concept and secured funding for the study visit to Cardiff with the assistance of A. Souza. All authors contributed to the study design. Testing and data collection were performed by C. Overkott. C. Overkott performed the data analysis with advice from both other authors. C. Overkott drafted the manuscript, and A. Souza and

C.C. Morey provided critical revisions. All authors approved the final version of the manuscript for submission.

Preliminary results from this project were presented at the 10th European Working Memory Symposium in September 2020.

Correspondence concerning this article should be addressed to Clara Overkott, Department of Psychology, University of Zurich, Binzmühlestrasse 14/22, 8050 Zürich, Switzerland, E-mail: claraoverkott@gmail.com or to Candice C. Morey, moreyc@cardiff.ac.uk.

Abstract

The capacity limitations of visual working memory may be bypassed by verbal labeling. In adults, labeling increases estimates of both quantity and quality of visual working memory. However, we do not know when children begin to use labeling and whether labeling similarly benefits visual memories of children under and over age 7. We assessed whether children benefit from prompted and spontaneous labeling opportunities, examining how labeling affects the storage of categorical (prototypical) and continuous (fine-grained) color information. Participants memorized colored candies for a continuous reproduction test either while remaining silent, labeling the colors aloud, or saying irrelevant syllables (discouraging verbal labeling). Mixture modeling confirmed that both categorical and continuous representations increased with age. Our labeling manipulation showed that spontaneous labeling increased with age. For the youngest children, prompted labeling especially boosted categorical memory, whereas labeling benefited categorical and continuous memory similarly in the older age groups.

Keywords: verbal labels, visual working memory, color categories, children, development, mixture modeling

Significance Statement

Though memory improves across childhood, it is not clear why. Verbal and visual memory are often considered separately, but in adults verbal and visual information interact. From around age 7, the frequency and fluency with which children use words to retain visual information improves, suggesting verbal labeling drives memory improvement. We tested memory for colors in children under 7 or 8 to 11 years old, manipulating whether participants were prompted to label the colors or whether labeling was inhibited. Labeling drastically increased the number of colors children under 7 could recall, and as labeling increased with age, the number of colors and precision with which they recalled the observed them improved. These findings reveal a developmental progression in how verbalization impacts memory for visual imagery.

The developing impact of verbal labels on visual memories in children

Though adult-level working memory for recently-observed visual information is known to be highly constrained (e.g., Luck & Vogel, 1997; Simons & Levin, 1997), it improves across childhood (e.g., Burnett Heyes et al., 2016; Cowan et al., 2010; Sarigiannidis et al., 2016; Simmering, 2012). Performance in visual working memory tasks improves steeply during childhood (Alloway et al., 2006; Burnett Heyes et al., 2012, 2016; Cowan et al., 2010, 2011; Cowan, 2016; Gathercole et al., 2004; Guillory et al., 2018; C. C. Morey, Hadley, et al., 2018; Sarigiannidis et al., 2016). Improvements across age have been observed in measurements of spatial memory (Alloway et al., 2006; Gathercole et al., 2004; Morey et al., 2018), complex visual working memory span (Swanson, 2017), visual change detection (Cowan et al., 2010, 2011; Shimi et al., 2014; Shimi & Scerif, 2015), and continuous reproduction of orientation (Burnett Heyes et al., 2012; 2016; Guillory et al., 2018; Sarigiannidis et al., 2016) and colors (Camos & Barrouillet, 2011; Shimi et al., 2014). Although the increasing trend is clear and robust, the reason for this improvement remains elusive: the interactivity of visual, verbal, and semantic processes means that there is a large candidate pool for *what* exactly develops that provokes improvement on visual working memory tasks.

Evidence from continuous reproduction tasks, in which participants select the precise value of some memorized feature from a continuous response set such as a color wheel (e.g., Brady et al., 2013) or orientation dial (e.g., Burnett Heyes et al., 2016) can be used to convincingly estimate the degree to which responses depend on categorical information (e.g., how close a color is to the canonical "red" of a British postbox), continuous information (e.g., a vivid representation of the precise hue observed), or sheer guesswork (Bae et al., 2015; Hardman et al., 2017). In adults, responses are informed by both categorical and continuous information, confirming that even recent visual memory depends on long-term knowledge (Bae et al., 2015; Donkin et al., 2015; Hardman et al., 2017; Pratte et al., 2017; Souza & Skóra, 2017). However, in children there is some ambiguity regarding which visual memory parameter improves (Burnett Heyes et al., 2012, 2016; Sarigiannidis et al., 2016) and, to the best of our knowledge, no study has attempted to measure categorical and continuous representations in children using the method of Hardman et al. We aim to confirm whether children's visual memory improves in terms of both quality and quantity, applying Hardman et al.'s procedure in order to take categorical biases into consideration, and additionally to discover how development of visual memory is influenced by verbal labeling.

For adults, generating a verbal label for a visual stimulus boosts both categorical memory and the precision of continuous representations (Forsberg et al., 2020; Overkott & Souza, 2021, 2021a; Souza et al., 2021). Souza and Skóra manipulated labeling opportunities, asking participants to either label to-be-remembered colors or repeat "bababa" aloud to inhibit labeling. They observed superior performance with labeling, with modeling showing effects of labeling on both categorical memory and continuous precision. Applying a verbal label did not merely replace any visual representation: the verbal label seemed to improve access to visual detail. These findings downplay the notion that verbal labels necessarily overwrite visual representations (e.g., Alogna et al., 2014), or even simply that verbal labels are maintained alongside a visual representation, which one might expect if participants were attempting to utilize all components of the classic multiple-component working memory model (Baddeley, 2012; Logie, 2011). Instead, verbal labels may facilitate integration with long-term knowledge, allowing visual information activated in long-term memory to be efficiently re-activated when needed (e.g., Hardman et al., 2017), resulting in overall improvements in how much detail from a recent episode can be precisely recalled. Verbal labeling may also serve to direct thought toward one representation and consequently away from others (e.g., Granato et al., 2020).

These findings suggest that verbal labeling may be particularly useful for boosting visual memory in individuals who typically struggle to remember, such as young children. Yet, it is unclear *if* and *how* labeling boosts children's visual memory. There is ample reason to speculate that spontaneous adoption of labeling differs in adults and children and, therefore, that development in fluency of labeling may account for some of the improvement in visual working memory tasks observed across childhood. We know that from age 5 to 10 years, children's tendency to spontaneously apply verbal labels to a serial picture memory task increases with age (Elliot et al., 2021; Flavell et al., 1966). This increase in verbalization in children could be driving improvement in visual working memory within this age range. Elliott et al. (in contrast to the original investigation of Flavell, et al.) found that most 5- and 6-year-olds (75% and 89% respectively) spontaneously verbalized during the memory tasks at least sometimes. For these children, overt labeling was most clearly associated with increased memory spans: the children who never labeled recalled less than the children who labeled at least sometimes. If the propensity for labeling is developing in children under 7, labeling may influence visual memory differently in younger children who are beginning to apply it than it does in older children and adults. Measuring and modeling components of visual memory with and without verbal labeling in children may therefore provide a clearer view of how visual memory develops. However, Elliott et al.'s stimuli (adapted from the original Flavell et al. materials) were images of nameable objects. For materials like these, recalling precise visual detail would have been unnecessary for completing the task. Indeed, we do not know from Elliott et al.'s results whether visual imagery was retained when verbal labeling was adopted. Some theories posit that labeling reflects a shift in preference toward phonological encoding and away from maintaining visual images (see Gathercole, 1998). If this interpretation is correct, then we could not infer from tasks using nameable objects whether verbal labels and visual images are both encoded. But by asking participants to remember a continuous value (e.g., like a color hue), whose verbal label alone is insufficient to generate a precise response, we can better pinpoint effects of verbal labeling on responses and learn how labeling and visual memory interact.

We seek to (a) establish whether children benefit from verbally labeling to-be-remembered continuous colors, (b) compare effects of labeling on mixture modeling in children younger and older than 7 with those of young adults, and (c) confirm whether representation of continuous as well as categorical information increases as children mature. In separate blocks of a continuous reproduction task, we manipulated whether color labeling was prompted (labeling condition), permitted covertly but unknown (silent condition), or inhibited with articulatory suppression (suppression condition). Comparing recall error in these conditions allows us to learn whether children benefit as much from opportunities to label as young adults do, and hints about how labeling effects may increase with development.

We applied the CatContModel of Hardman (2016) to our data to compare effects of labeling on parameters reflecting categorical as well as continuous visual memory representations. The CatContModel allows a more valid and realistic description of how both coarser categorical and detailed continuous representations influence memory responses through the assumption of categorical representational boundaries in the feature space. These boundaries may differ per age group, so we considered the content of the labels generated per group to inform assumptions about category boundaries. These features make the CatContModel ideal for assessing the potentially changing impact of labeling on visual memory and resolving contradictory claims about which aspects of visual memory are improving during childhood.

Methods

Participants

In total, 108 participants were tested. See Table 1 for a description of participants per age group. Adults were recruited from the undergraduate population of Cardiff University and children were recruited either from a local school or from the database maintained by the Cardiff University Centre for Human Developmental Sciences (CUCHDS). We targeted children between 5 and 7 years old and between 8 and 11 years old to compare participants likely to be transitioning toward labeling visual stimuli (the younger group) with older children who likely already spontaneously verbalize (cf. Elliott et al., 2021). Only fluent English-speaking participants reporting normal or corrected-to normal vision and no diagnosis of learning disability were eligible to take part in the study. The children were tested either at a local school or in the lab at CUCHDS, where the adults were tested. The study was approved by the Cardiff University School of Psychology's Ethics Committee (approval code EC.18.09.185339GR). Young adults gave written consent on their own behalf, while parents consented on behalf of their children. Children's assent to take part was assessed continuously via their behavior toward the task and researcher.

The research question, methods, and statistical hypotheses as well as possible outcomes of the reported experiment in this study were preregistered on the OSF under https://osf.io/5pd7n. The paradigm and analysis plan used in this paper follows the logic used in Souza and Skorà (2017). The experimental code, data and statistical analysis code can be found under https://osf.io/3wesd/. In our preregistration, we planned to test minimally 20 participants in each age group, but ideally, we would test more participants if time allowed. When negotiating to test children in the school, we committed to test all children in participating classes with consent who expressed interest. This resulted in testing nearly double the number of children planned; accordingly, we also doubled our sample of young adults. We did not recruit additional participants after beginning our data analysis. Two participants were excluded from the younger children group because they elected to stop the study without completing trials in the silent, labeling, and suppression conditions, leaving *N* = 32. Data from one younger child participant missing only the last block (always a second silent block) were included in the analyses. No other participants were excluded.

Table 1

Break-down of Participants' Ages and Gender per Age Group

	N	N female	Mean Age in Years (SD)
5 – 7-year-olds	32	18	6.85 (0.85)
8 – 11-year-olds	34	20	9.60 (1.00)
Adults	40	34	19.31 (2.10)

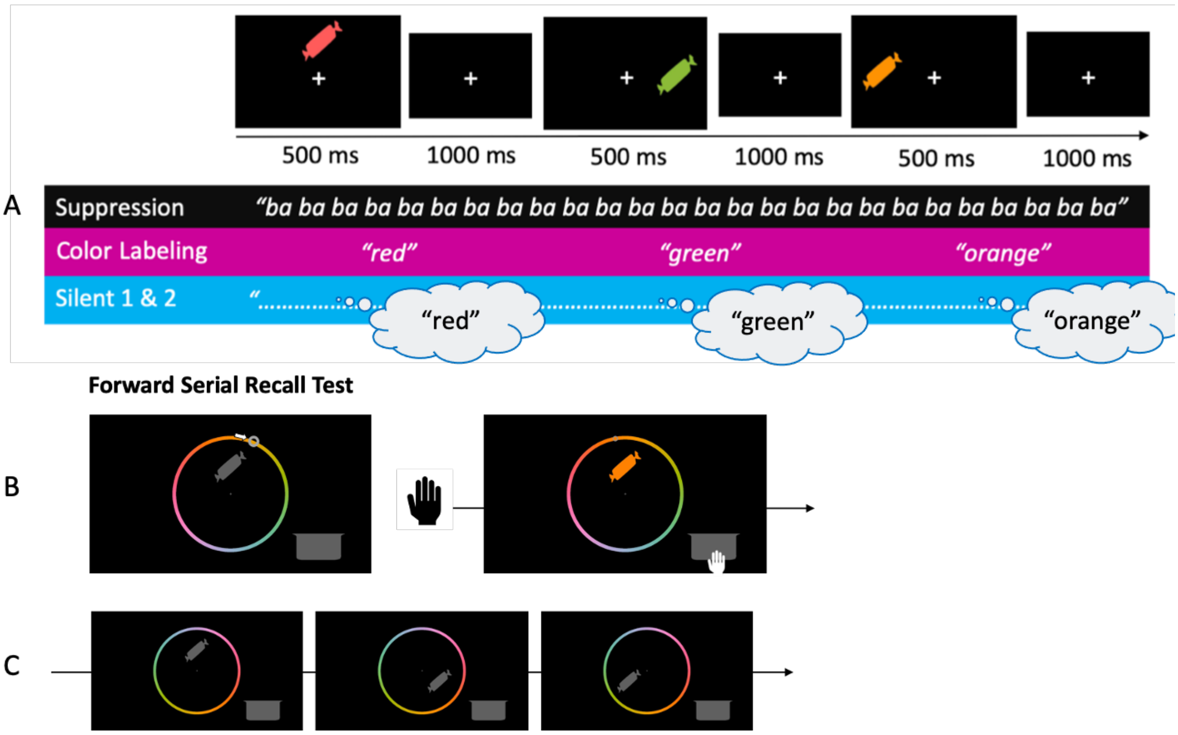
Materials & Procedure

All tasks were designed to operate via a touch-screen interface to ensure that even the youngest participants were able to respond in a natural manner. The experiment was programmed in MATLAB using the Psychophysics toolbox (Brainard, 1997; Pelli, 1997). Participants first completed a color matching task to train them how to use the color wheel. In this task, a colored candy (radius = 90 pixels), and a grey (RGB 96 96 96) candy were shown side-by-side on the screen against a black background (RGB 0 0 0). The task was to adjust the color of the grey candy (probe) to match the colored candy. Participants did this by moving a grey dot (RGB 150 150 150) with their finger on the laptop touchscreen along a color wheel that surrounded both candies. Once they started moving the dot along the wheel, the color of the grey candy changed to the color they were currently selecting on the wheel. Participants were instructed to adjust the probe candy until both candies were alike. When participants were satisfied with the chosen color, they clicked on a grey jar (length: 300 pixels, height: 150 pixels) that was presented in the right corner of the screen and the next candy appeared. The clip-art objects used in this study were taken from Sutterer and Awh (2016). Participants completed eight trials.

In the actual visual working memory task, each trial consisted of a study phase and a memory test phase. During the study phase, colored candies (memory items) were sequentially presented, each for 500 ms, followed by a 1000-ms inter-item blank interval (see Figure 1A). This timing was based on Cowan et al. (2010, 2011) to guarantee that children had enough time to encode and label each of the individual candies. The color of the candy was chosen randomly from 360 color values sampled from a color wheel created in CIELAB color space (L= 70, a= 20, b= 38, radius= 60). The colors in each trial were randomly selected with the only constraint that each color could only be used once per trial. To make the task requirements comparable between the children and young adults, we manipulated set size: both child groups were presented with three candies to remember and the young adults with four. We opted for this because previous research has shown that children can store, on average, half of the memory items young adults are able to remember (Cowan, 2016; Cowan et al., 2010, 2011). A set-size of four elements have been previously used to demonstrate the benefits of labeling in young adults (Forsberg et al., 2020; Souza and Skóra, 2017; Souza et al., 2021). For younger adults, lower set-sizes (1 or 2) generated ceiling performance that prevented the consistent observation of a labeling effect (Souza & Skóra, 2017). Having decided upon a set-size of 4 for adults, we could have halved the set-size for children. However, previous developmental research has shown that interventions to improve memory in children can be ineffective at a set-sizes of 2 but arise consistently at a set-size of 3 (Shimi & Scerif, 2017). Given that we tested two child age groups (for which capacity is assumed to be improving), we settled at using a set-size of three for the children and a set-size of four to the adults to maximize our chances of observing labeling benefits.

Figure 1

Example Trial of Study and Memory-Test Phase



Note. Panel A shows an illustration of the study phase with set-size three. Panel B shows the flow of the test phase of the experiment. Panel C shows the serial recall order for the test phase of each trial.

The candies were evenly spaced on a set of fixed positions on an imaginary circle (radius = 155 pixels) centered in the middle of the screen. Because of the differences in set size, preserving equal spacing between items resulted in distances between the fixed positions of the items of 120° between the candies for the children (forming a triangle on the screen) and 90° between the candies for the adults (forming a square). Our decisions to maximize spacing for each set size and present items one-at-a-time in constant locations were meant to minimize the likelihood of any inter-item interference. Furthermore, we considered that the even spacing of the memoranda across the screen would be less confusing for the children than to always have one unused location on the screen. In each trial, the candies were serially presented in these locations, starting from the top position, and proceeding in clockwise order.

Trials were organized into blocks with one of the following labeling manipulations: (a) *color labeling:* participants were asked to label the color of the candy aloud with whatever term they wanted (e.g., common terms such as red, green, blue; but also more unique terms such as orange-green, light blue, lavender, etc.), (b) silence: participants were asked to remain quiet during the study phase (and the supervising experimenter enforced this), but labeling could occur covertly, and (c) a suppression condition: participants were asked to repeat "bababa" aloud. The design of these conditions closely follows the ones used by Souza and Skóra (2017) and Forsberg et al. (2020) which have assessed labeling effects in young adults and young adults vs. older adults, respectively. These previous studies compared performance in these three types of labeling conditions under two assumptions. The first assumption was that articulatory suppression prevents verbalization but otherwise does not harm visual performance. The second one is that under silent conditions, participants can covertly label the memoranda, although they may do so less consistently than when overt labeling is required. This is in line with the observations in Experiment 4 of Souza and Skóra (2017). In this study, short presentation intervals that rendered labeling unlikely to happen generated similar performance under the silent, labeling, and articulatory suppression conditions. This shows that simply performing articulatory suppression continuously, even when there is not enough time to label, is inconsequential to visual working memory performance (see also Sense et al., 2017). Second, when there was sufficient time to label, performance in the articulatory suppression condition remained the same as in the short suppression condition, but performance in the silent and labeling conditions improved. This shows that participants benefitted from using this time for labeling. Yet, overt labeling produced somewhat better performance than silence (see also Forsberg et al., 2020). This is in line with the interpretation that labeling was occurring covertly in the silent conditions, but perhaps not as consistently. Here, we aimed to assess if: (a) overt labeling boosted performance compared to suppression in children as observed in young adults; (b) if silence produced better performance than suppression in line with the assumption that participants may be covertly labeling the memoranda; and (c) if overt labeling produced better performance than silence, suggesting that labeling was more consistently used when it was prompted.

Participants were asked to wear a headset throughout the sessions to record their verbalizations. Young adults completed all four blocks in one session, whereas children completed two blocks each across two sessions. To assess for possible carry-over effects of the manipulations, participants completed two blocks of silent trials, one always occurring first (*silent 1*) and one always last (*silent 2*). The order of the color labeling and suppression conditions (always second and third) were counterbalanced across participants. The silent 2 condition was used to test whether participants (especially children) would adopt spontaneous labeling after their experience in the prompted labeling condition.

During the memory test phase, all candies were recalled in serial order. First, a grey candy appeared at the top of the screen. As in the color matching task, a color wheel was shown surrounding the grey candy (Figure 1B). Participants were instructed to reproduce the color of the candy as accurately as possible by moving their finger along the color wheel on the touchscreen. They confirmed their response as practiced in the color matching task by touching the grey jar. Once they touched the jar, the remaining memory items were tested in clockwise order, in the same order they were presented (Figure 1C). Participants completed four practice trials before each condition, followed by 30 test trials, resulting in a total of 120 test trials. Given that participants recalled all memorized items, there were 90 responses (3 per trial) per condition for analysis for the two children groups, and 120 responses (4 per trial) for the young adult group.

Participants were instructed to start each trial in a self-paced manner by double-tapping the screen. They were reminded of the current labeling condition on each trial. After every third trial, participants received a visual feedback message that showed a jar that had more candies in it every time it was presented to keep participants (especially children) motivated. The feedback was independent of their ongoing performance. An experimenter was nearby (for children, in the same room) throughout the experiment to assist and encourage them, to prompt them if they did not perform the instructed labeling, or to remind them to remain silent in the silent blocks, as needed.

Data Analysis

Recall Performance

Like previous studies (Forsberg et al., 2020; Overkott & Souza, 2022; Souza & Skóra, 2017; Souza et al., 2021), our main dependent variable was recall error, defined as the angular distance in the color wheel between the participant’s response and the true color of the tested item. We submitted the data to Bayesian inferential analyses, which have several advantages over traditional *p*-value inference (Wetzels et al., 2011). One advantage is that the likelihood for an effect can be estimated in favor of the alternative or the null hypothesis. This likelihood can be reported with the use of Bayes Factors (BF), which gives the strength of evidence in favor of one model (i.e., the alternative) over another (i.e., the null) by calculating their marginal likelihood ratios given the data. The BF provides the factor by which our prior beliefs in the models under comparison should be updated in light of the data. A BF_{10} yields the likelihood of the alternative against the null, which can be reversed ($BF_{01} = 1/BF_{10}$) to express evidence in favor of the null. For example, a $BF_{10} = 3$ indicates that the alternative is three times more likely than the null hypothesis. A BF_{10} larger than 1 indicates evidence for an effect, and values lower than 1 provide evidence against an effect. BFs should be interpreted as a continuous index, however here we used the following guideline for classifying BFs. BFs larger than 3 were considered as providing nominal evidence in favor of an effect, whereas BFs larger than 10 were considered as strong evidence (Jeffreys, 1961). Our reported BFs were calculated in R (R Development Core Team, 2014) with default prior settings of the BayesFactor package (Morey & Rouder, 2015).

Categorical-Continuous Mixture Modeling

Similarly to previous studies assessing the labeling benefit (Forsberg et al., 2020; Overkott & Souza, 2022; Souza & Skóra, 2017, Souza et al., 2021), we also submitted participants' responses to the Bayesian hierarchical mixture model of Hardman et al. (2017) using the CatCont model package (Hardman, 2016) implemented in R. This categorical-continuous mixture model assumes that information about a studied color is either in memory (P^M), or not ($1– P^M$; guessing). Information in memory can be categorical knowledge ($1– P^O$), indicated by responses consisting of a color category (i.e., red, blue, green), or continuous knowledge (P^O), reflecting a representation that contains continuous information about the stimulus (i.e., light red hue). The latter representation is further defined by its imprecision (σ^O). The lower this value, the more precise the fine-grained response is. If no information is in memory, the model assumes that people guess either categorically (P^{AG}), reflected by guesses along the color categories, or continuously ($1– P^{AG}$) as uniformly distributed guesses.

In this model, every category has a mean and a standard deviation, which can be estimated freely by the model. Yet, the mean can be fixed when information about participants' color categories is known (see Souza & Skóra, 2017; Souza et al., 2021). We recorded participants' verbal responses to each presented color in the prompted labeling condition. We used this information to estimate the number of categories and their location (category means) for each group of participants, thereby fixing these values in the model. We note that this choice does not impact whether labeling benefits are observed, or the parameters in which they arise (see Souza & Skóra, 2017; Souza et al., 2021). We used the between-item variant of the model, as it has been reported to fit this type of data better (Hardman et al., 2017; Souza & Skóra, 2017). This model variant assumes that categorical and continuous information can be in memory at the same time, but at the point of response selection the response is based on only one source. Parameter values as well as the distributional probabilities were estimated with Markov Chain Monte Carlo (MCMC) sampling. We modeled the data of each age group separately.

Results

Verbal Labeling

Participants were required to wear headphones equipped with a microphone during the entire experiment to record their verbal responses during the study phase. Verbal responses in the prompted labeling condition were coded offline by the experimenter to assess the variety, number, and specificity of the color labels used by participants during the working memory task. The younger children used a total of 23 different labels to refer to the memoranda, the older children used a total of 33 labels, and the adults used 41 labels. For all age groups, seven common labels were most frequently used: red, orange, yellow, green, blue, purple, and pink. Uncommon labels, such as yellow-orange, teal or lavender were used in all age groups, but these individual labels occurred with very low frequency. We also classified responses as unintelligible when the labels were not understandable or when no label was provided. Figure 2A presents the overall frequency of these three label categories for each age group. This figure shows that the proportion of times common verbal labels were used was similar across age groups, with the younger children showing a larger tendency to remain silent during the trial than to use uncommon labels.

We evaluated how participants applied the seven commonly used color-labels to each of the 360 color-hues on our color wheel. The *x*-axis of Figures 2C to 2E show the color-hues of our color wheel in its unrotated form (i.e., 0° = red). We calculated the proportion of times each color label (e.g., “orange”) was used for each of the 360 color-hues, and this value was used as the *y*-axis of Figures 2C to 2E. A proportion of 1 indicates that 100% of the participants used the same color label when the color-hue indicated in the *x*-axis was memorized. Lower values are an indication of low agreement on the color label to be applied to that color-hue. For each of the seven color labels, we obtained a normal-like distribution along the circular space that rises when the memorized color-hue starts to resemble the label category (e.g., “orange”) and decreases again when the memorized color-hue starts to deviate from it. Take for example, Figure 2E that presents the adult labeling data. Each colored line in this figure represents one color label (“red", “orange", “yellow”, and so forth). Figure 2E shows seven bell-shaped distributions that represent the range of color-hues for which each color-label was applied. Figures 2C and 2D show the same information for the two children groups. Overall, these panels show that all age groups had the same intuitions about applying the same labels to the same color-hues. Younger children agreed on the color labels (hitting 1.0), but there was more variation and noise in between the color categories in comparison to the other age groups, which showed progressively less variability.

To confirm the similarity in color labeling across groups, we estimated the mean and the standard deviation of each color-label category for each age group (see Figure 2B). This measure indicates the position of the color category and its spread over the wheel. This was done as follows: the frequency of use of a given color label as a function of the studied color-hue on the wheel resembles a normal distribution over the color space (see Figures 2C-E). A von Mises distribution (normal distribution for the circular space) was fitted to these distributions to estimate the mean and the imprecision (standard deviation) of each color category. Figure 2B shows that the mean for each color category is alike for all three age groups indicating that children and younger adults have similar color categories for this color space.

Color Matching

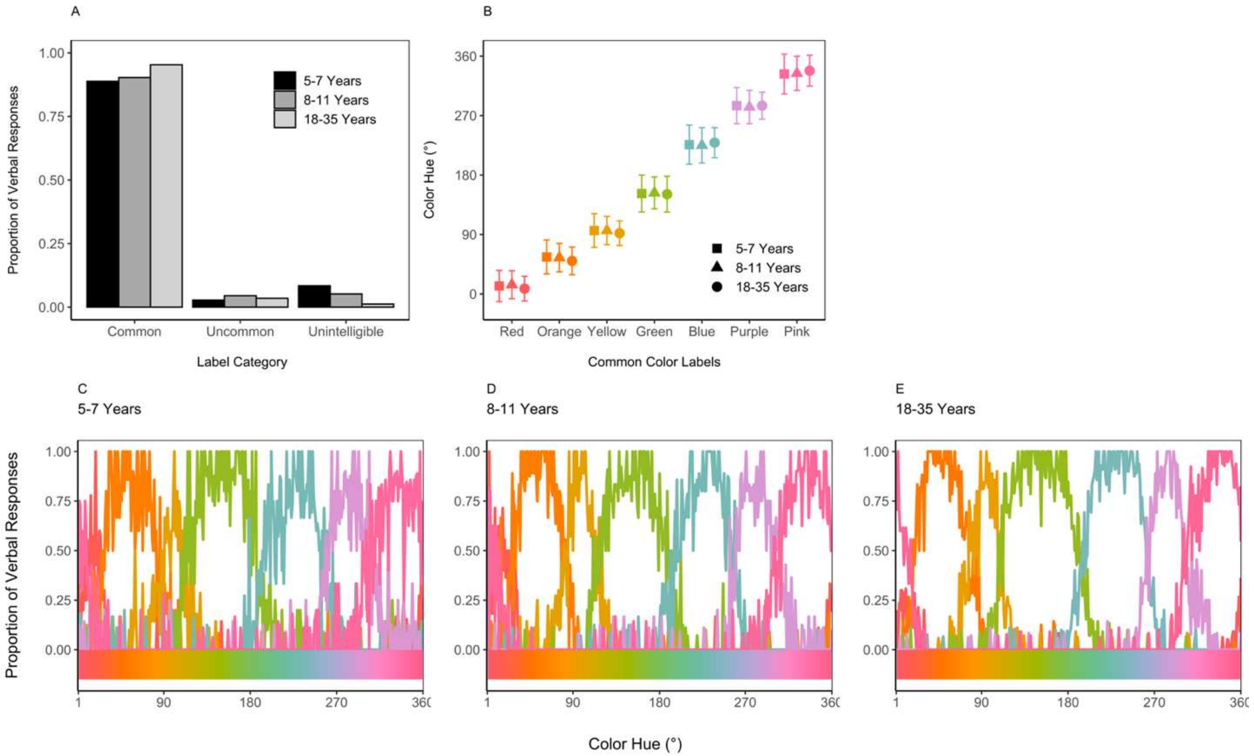
An analysis of the performance in the color matching task revealed that our participants did understand the experimental procedure and the use of the color wheel. The mean error in the color matching task was only 8.54 for the younger children, 4.76 for the older children and 3.70 for the adults, respectively. This is similar to previous developmental studies (Burnett Heyes et al., 2012; Shimi & Scerif, 2021).

Recall Performance

Figure 3 shows the mean recall error in each labeling condition separately for each age group. Although adults memorized four items whereas children memorized only three, recall error was lowest for adults in all conditions, which confirms our decision to use fewer items for children to avoid poor performance and discouragement. As expected from an increase in visual working memory capacity with age, performance improved with age: the younger children's recall error was higher than the older children's, whose recall error was higher than the adults. Note, however, that comparisons on the recall measure alone are confounded by differences in set-size between children and adults. Developmental changes in working memory capacity will be more directly addressed in discussing the capacity measure K.

Figure 2

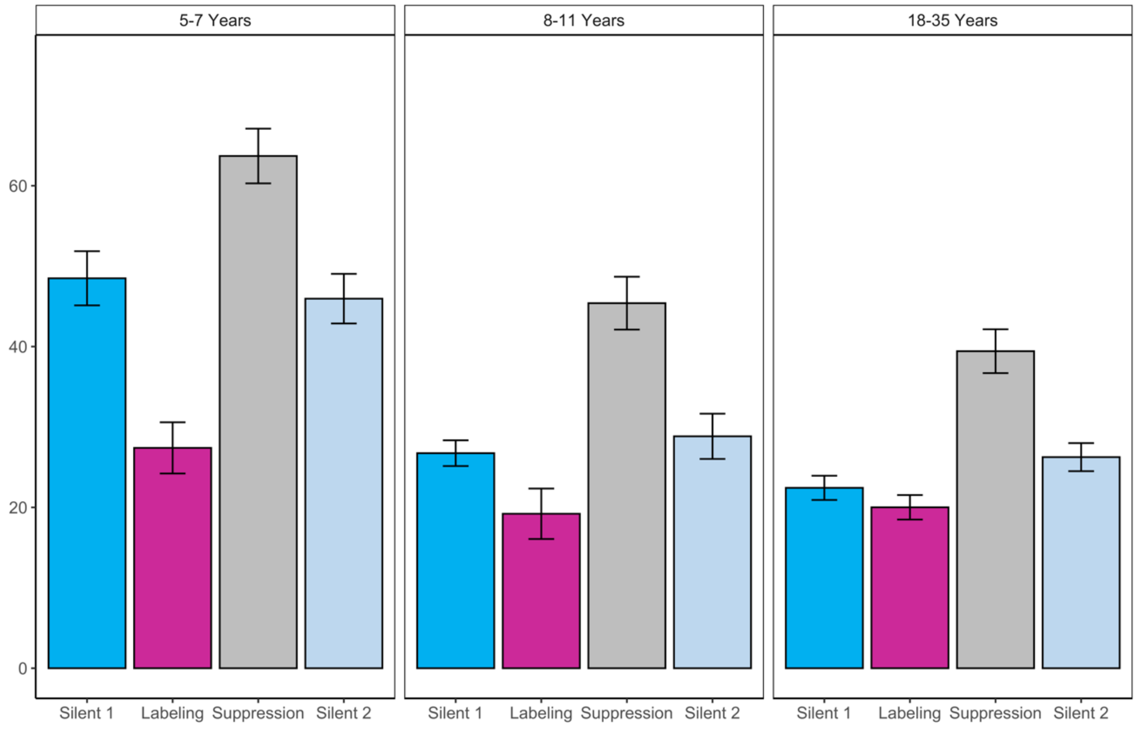
Analysis of Color Labels used by the Participants for all Age Groups



Note. Panel A represents the proportion of verbal responses in the labeling condition grouped by common, uncommon, and unintelligible labels for each age group. Panel B shows the mean color for a given color label and its standard deviation for each age group. These parameters were obtained by fitting a von Mises distribution on the distribution of verbal responses over the color space. Panels C, D, and E show the frequency with which each verbal label (represented by the different colored lines) was applied to each color on the wheel (x-axis) in each of the three age groups. Each verbal label (e.g., orange) is represented by its prototypical wheel color. A proportion of 1.00 indicates that the color term was used by all participants when the color hue depicted on the x-axis was presented. The lower the proportion, the less often the color term was used to refer to that given color on the color space.

Figure 3

Mean Recall Error (in Degrees) in Each Labeling Condition Presented Separately for Each Age Group



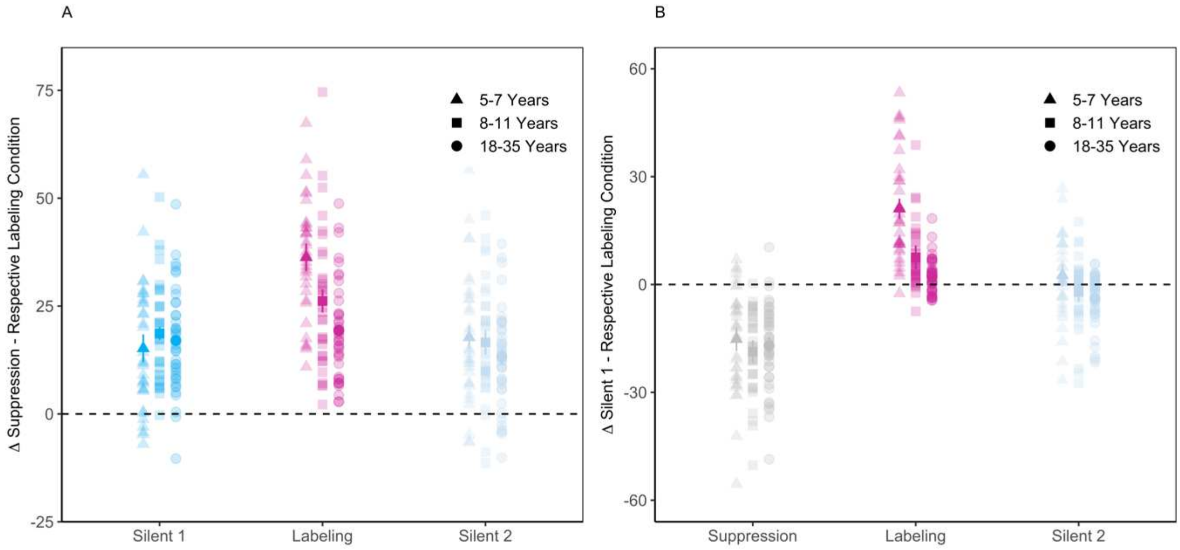
Note. Error bars represent the 95% within-subjects confidence interval.

Critical to our research question, Figure 4A shows the absolute difference between the suppression condition and all other conditions, and Figure 4B between the silent 1 condition and each other condition. The zero line represents no difference in performance between conditions, whereas values higher than zero represent an increase in performance compared to the reference condition, or a decrease in performance compared to the reference condition.

As shown in Figures 3 and 4, when prompted to label the color aloud (aka labeling condition), performance improved in comparison to both suppression and the silent 1 condition, respectively, for all age groups. Similarly, in Figure 3 performance in both silent conditions was better than in the suppression condition, consistent with the hypothesis that participants engaged in spontaneous labeling in the silent condition. There was no performance improvement from the silent 1 to the silent 2 condition, suggesting no carry-over from exposure to the prompted labeling condition (and hence no change in the willingness of adopting a labeling strategy) in the second silent block, even for the younger children (see Figure 3). With regards to the silent conditions, we can see a clear difference among the age groups: performance in the silent condition was not as good as in the labeling condition for the younger children, but this difference became progressively smaller for the older groups. To simplify the analysis, from here on only the silent 1 condition will be considered for analysis, given that there were no systematic differences between the two silent conditions (see Supplementary Analysis on OSF https://osf.io/ex2hd for full analysis).

Figure 4

Absolute Difference in Recall Error Between the Suppression Condition and each Labeling Condition Presented Separately for Each Age Group



Note. Error bars represent the 95% within-subjects confidence interval. Smaller, paler symbols depict the average relative difference in recall performance between conditions of individual participants.

Additionally, in Figure 4A labeling benefitted performance in contrast to the silent 1 condition and this benefit seems to decline from the young to the older children and adults. This is in line with the assumption that though children engaged in verbal labeling, labeling was more likely in the older children group.

As mentioned in the preregistration, we performed a Bayesian ANOVA (hereafter BANOVA) consisting of a between-subjects factor (age group: 5-7 vs. 8-11 vs. 18-35) and a within-subjects factor (labeling condition: silent 1 vs. color labeling vs. suppression), presented in Table 2. The best model included both main effects and their interaction, with the inclusion of the interaction term being overwhelmingly supported. This result is consistent with the visual trends that overall performance was best for the adults compared to the children, and that prompted labeling yielded the best performance, followed by silence, and then suppression. The interaction reflects the fact that the effect of labeling varied as a function of age group, supporting the notion that there is a developmental trend for the utility of labeling.

We further preregistered analyses of possible outcomes to answer our first two research questions. First, if articulatory suppression impairs ability to covertly label, the comparison between the suppression and the labeling condition allowed us to estimate the effect of verbal labeling. We aimed to assess whether the labeling benefit changed in size across development. To assess this, we ran the same BANOVA as before, but the condition factor included only the levels of suppression and labeling (see Table 2). All age groups benefitted from prompted labeling and this benefit was largest for the youngest children (see Figure 4). Thus, the effect of prompted verbal labeling increased across development, thereby answering our first research question.

Table 2

Bayes Factor (BF) of Models with Different Fixed Effects Over the Null and BF favoring the Best Model Over the Model Specified in Each Row (BF_{Best}/BF_{Mrow})

Best Model Over the Model Specified in Each Row (BF_{Best}/BF_{Mrow})

Comparison	Model n°	Included Fixed Effects			BF_{10}	BF_{Best}/BF_{Mrow}
		Age Group	Condition	Age x Cond.		
Full	**1**	✓	✓	✓	**2.69×10^{62}**	**1**
	2	✓	✓	---	6.14×10^{55}	4.38×10^{6}
	3	✓	---	---	8.70×10^{7}	3.09×10^{54}
	4	---	✓	---	1.62×10^{47}	1.66×10^{15}
Suppression - Labeling	**1**	✓	✓	✓	**3.19×10^{43}**	**1**
	2	✓	✓	---	9.32×10^{39}	3.42×10^{3}
	3	✓	---	---	1.60×10^{3}	1.99×10^{40}
	4	---	✓	---	9.44×10^{33}	3.38×10^{9}
Suppression - Silent 1	1	✓	✓	✓	2.99×10^{30}	7.06
	2	✓	✓	**---**	**2.11×10^{31}**	**1**
	3	✓	---	---	2.28×10^{9}	9.27×10^{21}
	4	---	✓	---	7.48×10^{21}	2.83×10^{9}
Labeling - Silent 1	**1**	✓	✓	✓	**1.08×10^{25}**	**1**
	2	✓	✓	---	8.83×10^{16}	1.22×10^{8}
	3	✓	---	---	5.19×10^{7}	2.08×10^{17}
	4	---	✓	---	1.20×10^{9}	8.97×10^{15}

Note. ✓ = effect included in the model. Best model is printed in bold. Best model = model with higher BF over the Null. Comparable analyses were conducted adding individual color-matching performance as a covariate, which yielded the same inferential outcomes. These analyses can be found in our online supplement (https://osf.io/ex2hd).

Second, the comparison between the silent and suppression conditions allowed us to estimate whether participants might be engaging in spontaneous verbal labeling. For this, we ran a BANOVA including the levels suppression and silent 1 for the condition factor. The best model included only the two main effects (Table 2), and there was substantial evidence against including the interaction term in the model, indicating that the effect of suppression on color reproduction was similar for all age groups (see Figure 4)[1].

Additionally, we carried out an unplanned comparison of the labeling and silent condition, as an additional check of the prompted labeling benefit. In adults, there is ample evidence that articulatory suppression requires little attention and does not much impair concurrent visual memory tasks (Sense et al., 2017; Souza & Skóra, 2017 Experiment 4). However, it is possible that for children the instruction to continuously say meaningless syllables demanded attention and affected visual working memory because of that burden. Comparing the prompted labeling and silent conditions provides an extra check that labeling itself benefited visual memory, rather than suppression impairing it. The best model included both main effects and their interaction, just as the suppression-labeling comparison did. Here, all groups benefited from labeling, but the youngest children benefited most.

Categorical-Continuous Memory

We performed mixture modeling to estimate how continuous and categorical representations changed with age and across the labeling conditions. To get a first glimpse of the distribution of categorical and continuous responses, Figure 5 presents scatterplots relating the studied color against the reported color for each labeling condition and age group. In the scatterplots, we can detect a mixture of categorical responses that cluster around certain color values (i.e., red, blue, green) in a stepwise manner along the diagonal. Then,

---

[1] The full analysis on the preregistered comparison of first and second silent block can be found in the Supplementary Analysis under OSF: https://osf.io/ex2hd/.

continuous responses that align along the vertical line, with the dispersion around this line reflecting the precision of the responses. Guessing behavior is reflected by randomly distributed points. Across all labeling conditions, we can detect categorical clusters along the vertical line. The suppression condition contains more disperse points, and the labeling condition has the densest diagonal line. This pattern can be observed for all age groups. The mixture model we applied to the data has parameters to estimate the contribution of each of these sources to responses, which are not captured by other types of mixture models that do not include parameters to account for categorical responses. It also allow us to separately assess benefits of labeling to the retention of only categorical responses from boosts to the storage of continuous representations.

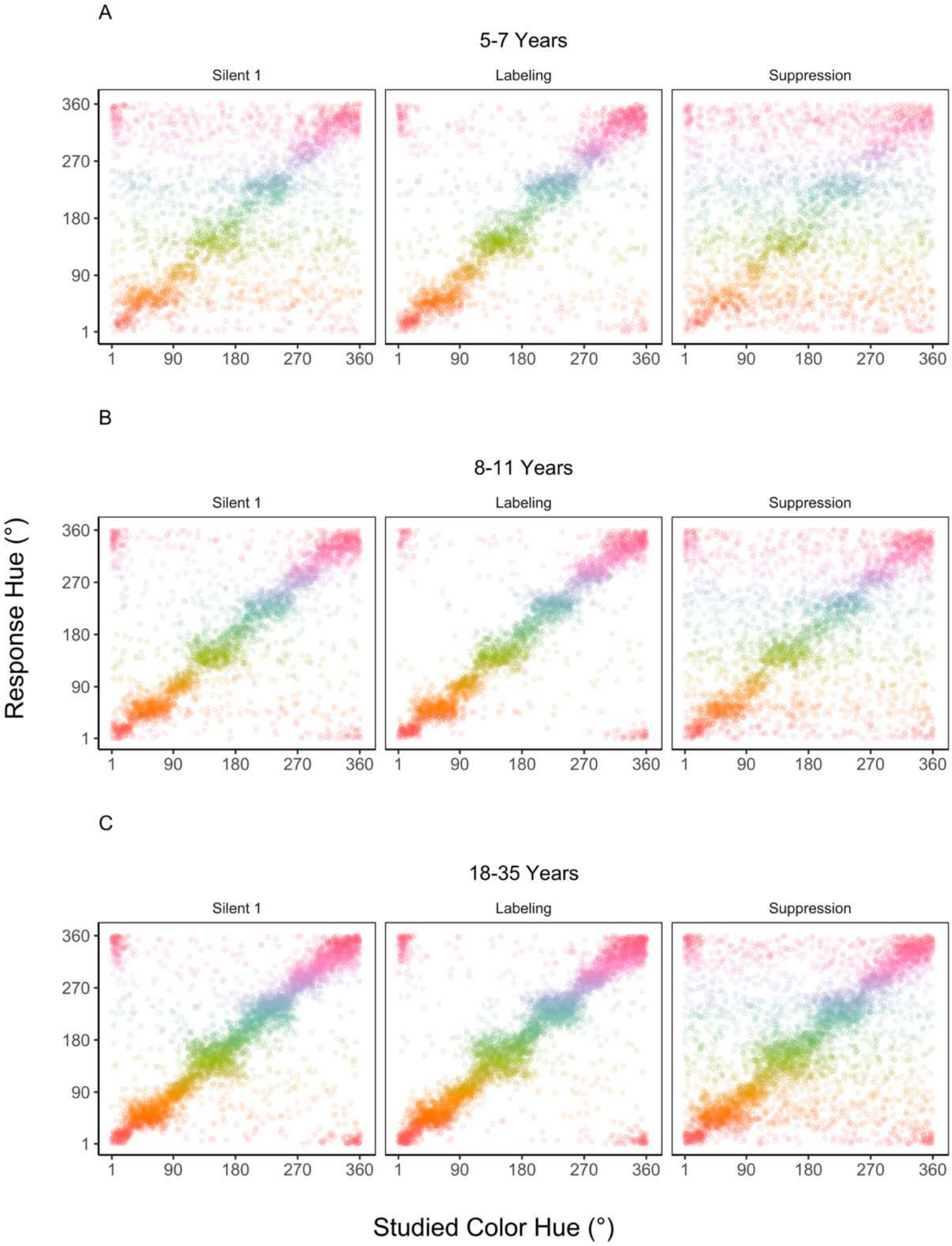
We modeled the data of each age group with the CatContModel package (Hardman, 2016). Each model contained 10,000 iterations, of which the first 2,000 were regarded as burn-in. In the model, we fixed the number of color categories to seven and we informed the model about the category means obtained from the verbal output data of each age group (as in Souza & Skóra, 2017; Souza et al., 2021). The color categories were assumed to be the same for all participants within an age group.

This mixture model allows for condition effects on three parameters: probability that responses are informed by memory (P^M), probability that memory responses are informed by continuous representations (P^O), and the imprecision of the continuous representations (σ^O). We assessed the group-level posterior estimates of these parameters in a model in which we allowed for an effect of condition on all parameters (see Table 3). To calculate the probability of categorical memory, we need to multiply P^M with $1– P^O$, thereby reflecting the amount of memory responses based on categorical information in memory. To calculate continuous memory, we multiply P^M with P^O, which reflects the information in memory that contains fine-grained detail about the visual feature studied. To exemplify, let us assume that the model estimates $P^M = 0.7$ and $P^O = 0.4$ in a given condition. This indicates that 70% of the responses were informed by memory, and the remaining 30% reflected guessing. To know how much of the memory responses are based on categorical vs. continuous information, we need to conditionalize P^M by P^O. The proportion of categorical memory equals $0.7 \times (1 - 0.4) = 0.42$ (42%) as opposed to continuous memory of $0.7 \times 0.4 = 0.28$ (28%). The continuous imprecision parameter (σ^O) was used as outputted by the model. A more detailed analysis on the posterior effects of the model, including the silent 2 condition can be found in the supplementary analysis on the OSF https://osf.io/ex2hd.

Table 3 shows that for the younger children group, when comparing the 95% highest density intervals (HDIs) for the labeling condition with the suppression condition, there was a clear labeling benefit for both categorical and continuous memory. This is reflected in the non-overlapping HDIs for these conditions. Likewise, both categorical and continuous memory improved when comparing the silent with the labeling condition (though the effect is less clear for continuous memory, as there the 95% HDIs slightly overlapped). For the older children group, there was a labeling benefit for both the silent and the labeling condition in categorical memory in comparison to the suppression condition, and a clear benefit of labeling over suppression for continuous memory. Similar modeling results can be found for the adult group. These results suggest that labeling benefits both categorical and continuous memory in all age groups, but that benefits of overt labeling instructions are strongest for the youngest children. For all age groups, there was no credible effect of labeling on continuous imprecision, which reflects the variability of the continuous representation maintained.

Figure 5

Scatterplots of Study-Response Distribution for all Age Groups



Note. Panels A shows the response color as a function of the studied color for the four labeling conditions for the 5-7-year-olds, Panel B for the 8-11-year-olds and Panel C for the 18-35-year-olds.

Table 3

Posterior Means and Highest Density Intervals (HDI) for all Age Groups

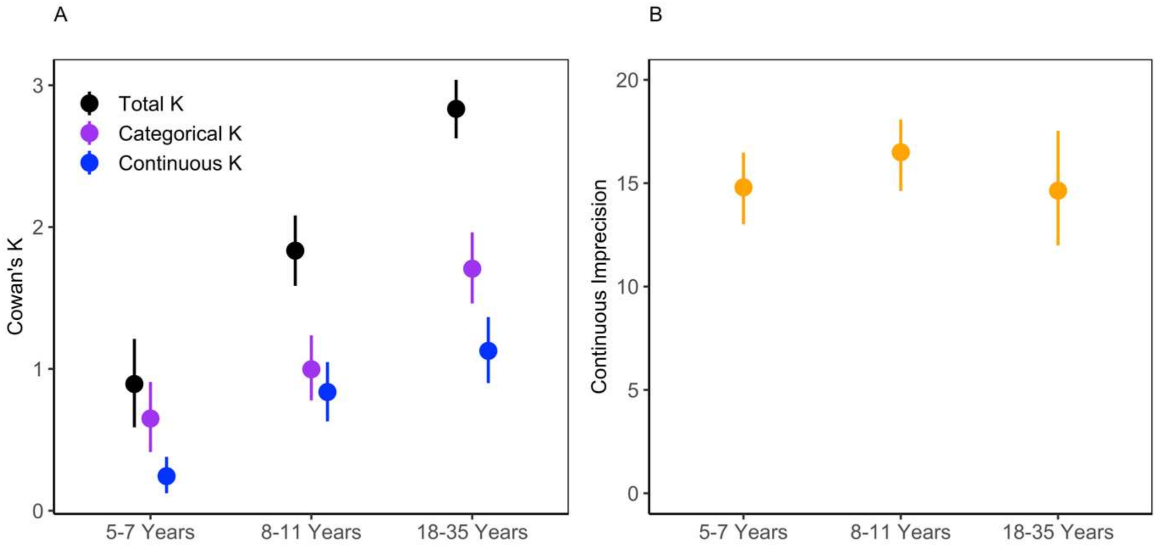
	Categorical		Continuous		Cont. Imprecision	
Age Group + Condition	Mean	95 % HDI	Mean	95 % HDI	Mean	95 % HDI
5-7 Years						
Silent 1	0.40	[0.30-0.51]	0.18	[0.11-0.26]	14.47	[12.30-16.90]
Labeling	0.63	[0.54-0.72]	0.30	[0.21-0.38]	13.65	[12.17-15.19]
Suppression	0.21	[0.14-0.30]	0.08	[0.04-0.13]	14.64	[11.99-17.54]
8-11 Years						
Silent 1	0.50	[0.42-0.57]	0.39	[0.32-0.47]	14.18	[12.84-15.43]
Labeling	0.54	[0.46-0.62]	0.44	[0.36-0.51]	14.05	[12.91-15.22]
Suppression	0.33	[0.26-0.41]	0.28	[0.21-0.35]	16.50	[14.62-18.08]
18-35 Years						
Silent 1	0.45	[0.39-0.51]	0.47	[0.41-0.53]	12.82	[11.97-13.75]
Labeling	0.54	[0.48-0.60]	0.42	[0.36-0.49]	12.95	[12.02-13.91]
Suppression	0.43	[0.37-0.49]	0.28	[0.22-0.34]	14.80	[13.02-16.48]

To assess the development of storage capacity for categorical and continuous memory across development, we further calculated Cowan's K (Cowan, 2001) for each age group in the suppression condition. The suppression condition affords the purest estimate of developmental trends in visual working memory because it reduces reliance on verbal labeling. The advantage of this capacity K measure is that it transforms the probability that information is in memory in relation to the different set sizes memorized by the children and adults, which allows a more interpretable comparison of performance across age groups.

Based on the estimated modeling parameters (Table 3), we calculated total K, categorical K, and continuous K. Total K - the sum of categorical and continuous representations - was estimated by multiplying $P^M \times$ set-size. Categorical K was calculated by $P^M \times (1-P^O) \times$ set-size. Continuous K resulted in $P^M \times P^O \times$ set-size. Each of the K estimates (total, categorical, and continuous) was calculated separately for each age group. Figure 6A shows that there is a steep increase in total K across development. The pattern of categorical and continuous K varies. Categorical K improves more drastically from older children to young adults, which can be detected by the non-overlapping ranges of highest density intervals. In contrast, continuous K increases most between the two child groups. Continuous memory imprecision (Figure 6B) did not indicate any credible change with age.

Figure 6

Cowan's K (Panel A) and continuous imprecision (Panel B) for the Suppression Condition across all Age Groups.



Note. The dots depict the mean and the error bars the 95% HDI.

Discussion

Can children benefit from verbal labeling and how does this affect the storage of categorical and continuous visual information? To assess this, we administered a continuous color reproduction task to children aged 5-7 years, children aged 8-11 years, and young adults, and applied mixture modeling using a procedure that accounted for the influence of color-category boundaries. Participants were asked to either (a) say the colors aloud to assess prompted labeling or (b) remain silent to assess whether they would benefit from the free time to covertly (and spontaneously) adopt labeling. These conditions were contrasted with an articulatory suppression condition intended to inhibit verbal labeling and provide a comparatively purer measure of visual working memory. We also contrasted the prompted labeling condition to the silence condition to see if benefits of labeling would remain even when any potential disruption introduced by the articulatory suppression is removed.

First, both younger and older children benefitted from prompted labeling (whether compared against suppression or silent conditions), which suggests that children under seven can use verbal labeling when instructed to do so (Cowan et al., 2011; Henry et al., 2012; Jarrold & Citroën, 2013; Keeney et al., 1967; Vales & Smith, 2015). Labeling may help prevent attentional lapses. Moreover, this instructed labeling benefit was highest for the younger children group – consistent with the advantages Elliott et al. (2021) observed when children were instructed to name picture stimuli – indicating that labeling can narrow the age gap in visual working memory tasks.

Second, performance in the silent condition was better than in the suppression condition for all ages. The magnitude of improvement in the silence condition against the suppression condition was similar for all age groups. However, the appearance of similar outcomes does not necessarily mean that the same underlying processes produced these outcomes. We can think of two reasons why suppression yields poorer performance than silence: engaging in suppression may impose an additional dual-task burden, or may inhibit possibilities to label. Recall may be better in silence than under suppression because silence affords labeling or because silence imposes no dual-task burden. Additionally, the underlying reason for the effect of suppression may differ per age age group. We know that suppression does not much affect visual working memory in adults (Sense, et al., 2017; Souza & Skóra, 2017) but engaging in suppression is possibly a greater burden for younger children. However, if suppression were more demanding for the children than for adults, we would expect performance in the suppression condition to be more impaired for the children than adults compared to silence. This was not the case. Furthermore, the argument that suppression would be more impairing for the children is based on the notion that speaking the syllables would serve as a dual task. However, under this logic, dual-tasking was also required in the labeling condition: participants had to find and apply the correct label to the colors and say them aloud within a short-time limit (1500 ms). If children disproportionately struggle with dual-tasking, labeling could even generate a cost. Labeling would compete with memorizing the precise color hue, leading to less precise responses. This was not the case here, as younger children benefited more from labeling than older children and adults.

Given these arguments, we contend that a simpler explanation of our findings is that articulatory suppression had a main effect for all age groups: inhibit labeling. When labeling was not inhibited, participants could covertly label. The extent they did so is reflected by the comparison of the silent condition to the labeling condition: this benefit of prompted labeling was largest for the younger children. Overall, these findings suggest that even children under seven may have benefitted from spontaneous labeling opportunities, suggesting that children younger than seven can use proactive verbal strategies to boost visual working memory maintenance. Yet, the fact that the effect of prompted labeling was largest for the younger children highlights that covert labeling may have been less frequent in the younger children. They needed the prompt from the experimenter to apply this strategy in all trials and reap its benefits. Altogether, this pattern is consistent with the hypothesis that spontaneous labeling increases with development. This development cannot be explained by assuming labeling is more costly in younger children, given that when labeling was consistently used, memory improved substantially. This pattern is remarkably consistent with the rates of overt, spontaneous labeling Elliott et al. (2021) documented in 5–10-year-old children: though older children labelled more frequently and consistently, a substantial portion of Elliott et al.'s 5- and 6- year-old children were observed spontaneously labeling picture stimuli. Here, we found consistent evidence using a different task, which bolsters the contention that children younger than 7 often adopt and benefit from verbalization, although they may apply it less consistently. Altogether, these findings are consistent with the assumption of a transition towards more proactive mnemonic strategies as children approach age 7 (Morey, Hadley, et al., 2018; Morey, Mareva, et al., 2018).

Our results also provide important insights into developmental changes in visual working memory capacity. In the suppression condition, we observed that both categorical and continuous representations increased with age, although we did not observe any credible differences in terms of continuous imprecision. Previous studies showed improvements in memory imprecision (Burnett Heyes et al., 2012, 2016) or in probability of recall (Sarigiannidis et al., 2016) as children aged, but neither considered categorical responses, which may subserve less precise memories. This is because categorical responses are, by definition, less precise responses. Hence more reliance on categorical representations would be accommodated by traditional mixture models as less precise memories. We introduced some important elements to better build on previous work and resolve these discrepancies. First, our sample included younger children than previously tested, which afforded greater scope for detecting developmental changes. Second, we applied the CatContModel (Hardman et al., 2017) for the first time in children, considering color category boundaries specific to each age group. With this we could observe that younger children stored very few continuous representations, and while this estimate increased for the older children, categorical memory remained comparable. The contrast between older children and adults revealed a further gain in categorical memory with age. Separating categorical and continuous representations allowed us to observe that the imprecision of continuous representations was similar across age groups.

Finally, instructed labeling was helpful for the retention of categorical and continuous memory in all age groups. In other words, the probability of storing both categorical and continuous representations was larger in the labeling compared to suppression (and silent condition) for all age groups. The only notable developmental effect was that prompted labeling was especially helpful for maintaining categorical information in the youngest children, revealing a difference in how labeling uniquely benefits older children and adults' visual memory. Why is the prompted labeling benefit in children 7 and younger most strongly tied to a categorical increase? One possibility is that the retention of continuous memory is costly, and the retention of many continuous representations exceeds the capacity limitations of younger children. Retaining less costly categorical information may allow the youngest children to overcome these limits.

At a first glance, the finding that prompted labeling for young children particularly benefitted categorical memory seems consistent with the dual-trace hypothesis, assuming that a verbal trace of the label and a visual trace of the input are both maintained. One might be tempted to further argue that these representations are held in distinct working memory stores (Baddeley, 2012; Logie, 2011). However, this cannot explain the whole developmental pattern we observed. Firstly, this hypothesis predicts only a change in categorical and not in continuous memory with labeling, which is not what we found in any age group. Moreover, a proponent of this position would need to explain why the impact of labeling shifts from boosting categorical representations to boosting continuous ones from age 7 onward; if one supposes that younger children store these categorical representations in a distinct buffer, what happens to it later? Finally, our results are not in line with the verbal recoding hypothesis in which a verbal categorial representation is saved instead of a visual representation, because this predicts a decrease in continuous memory that we did not observe.

Instead, our finding of increased continuous memory in all age groups when labeling supports the hypothesis that labeling activates categorical knowledge in long-term memory (Overkott & Souza, 2020, 2021a; Souza et al., 2020; Souza & Skóra, 2017). The activation of a category in long-term memory through the verbal label may facilitate data compression. The label acts as a reference, and the visual input can be encoded or consolidated in relation to this reference, while removing redundancy. This would allow more information to be held in long-term memory, freeing up limited working memory capacity. Moreover, it may protect the current representation in working memory from interference.

In conclusion, our findings confirm that visual working memory performance improves across childhood in terms of both categorical and continuous representations. Our labeling manipulation revealed a possible reason behind this increase: the verbal label is beneficial to memorization of visual imagery, and the spontaneous use of verbal labeling may increase with development, which may be one of the substantial drivers of developmental improvement to visual working memory capacity.

References

Alloway, T. P., Gathercole, S. E., & Pickering, S. J. (2006). Verbal and visuospatial short-term and working memory in children: Are they separable? *Child Development*, *77*(6), 1698–1716.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Bornstein, B., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., & others. (2014). Contribution to Alonga et al (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578.

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29.

Bae, G.-Y., Olkkonen, M., Allred, S. R., & Flombaum, J. I. (2015). Why some colors appear more memorable than others: A model combining categories and particulars in color working memory. *Journal of Experimental Psychology: General*, *144*(4), 744.

Brady, T. F., Konkle, T., Gill, J., Oliva, A., & Alvarez, G. A. (2013). Visual long-term memory has the same limit on fidelity as visual working memory. *Psychological Science*, *24*(6), 981–990.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897X00357

Burnett Heyes, S., Zokaei, N., & Husain, M. (2016). Longitudinal development of visual working memory precision in childhood and early adolescence. *Cognitive Development*, *39*, 36–44. https://doi.org/10.1016/j.cogdev.2016.03.004

Burnett Heyes, S., Zokaei, N., van der Staaij, I., Bays, P. M., & Husain, M. (2012). Development of visual working memory precision in childhood. *Developmental Science*, *15*(4), 528–539.

Camos, V., & Barrouillet, P. (2011). Developmental change in working memory strategies: From passive maintenance to active refreshing. *Developmental Psychology*, *47*(3), 898.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114-185.

Cowan, N. (2016). Working Memory Maturation: Can We Get at the Essence of Cognitive Growth? *Perspectives on Psychological Science*, *11*(2), 239–264. https://doi.org/10.1177/1745691615621279

Cowan, N., AuBuchon, A. M., Gilchrist, A. L., Ricker, T. J., & Saults, J. S. (2011). Age differences in visual working memory capacity: Not based on encoding limitations: Differences in working memory capacity. *Developmental Science*, *14*(5), 1066–1074. https://doi.org/10.1111/j.1467-7687.2011.01060.x

Cowan, N., Morey, C. C., AuBuchon, A. M., Zwilling, C. E., & Gilchrist, A. L. (2010). Seven-year-olds allocate attention like adults unless working memory is overloaded: Capacity and attention allocation. *Developmental Science*, *13*(1), 120–133. https://doi.org/10.1111/j.1467-7687.2009.00864.x

Donkin, C., Nosofsky, R., Gold, J., & Shiffrin, R. (2015). Verbal labeling, gradual decay, and sudden death in visual short-term memory. *Psychonomic Bulletin & Review*, *22*(1), 170–178.

Elliott, E., Morey, C. C., AuBuchon, A., Cowan, N., Jarrold, C. R., Adams, E., Attwood, M., Bayram, B., Beeler-Duden, S., & Blakstvedt, T. (2021). Multi-lab direct replication of Flavell, Beach, & Chinsky (1966): Spontaneous verbal rehearsal in a memory task as a function of age. *Advances in Methods and Practices in Psychological Science*.

Flavell, J. H., Beach, D. R., & Chinsky, J. M. (1966). Spontaneous verbal rehearsal in a memory task as a function of age. *Child Development*, 283–299.

Forsberg, A., Johnson, W., & Logie, R. H. (2020). Cognitive aging and verbal labeling in continuous visual memory. *Memory & Cognition*.

Gathercole, S. E. (1998). The development of memory. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(1), 3–27.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177.

Granato, G., Borghi, A. M., & Baldassarre, G. (2020). A computational model of language functions in flexible goal-directed behaviour. *Scientific Reports*, *10*(1), 1–13.

Guillory, S. B., Gliga, T., & Kaldy, Z. (2018). Quantifying attentional effects on the fidelity and biases of visual working memory in young children. *Journal of Experimental Child Psychology*, *167*, 146–161. https://doi.org/10.1016/j.jecp.2017.10.005

Hardman, K. O. (2016). *CatContModel: Categorical and Continuous Working Memory Models for Delayed Estimation Tasks* (0.7.5) [Computer Software]. https://github.com/hardmanko/CatContModel/releases/tag/v0.7.5

Hardman, K. O., Vergauwe, E., & Ricker, T. J. (2017). Categorical working memory representations are used in delayed estimation of continuous colors. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 30–54. https://doi.org/10.1037/xhp0000290

Henry, L. A., Messer, D., Luger-Klein, S., & Crane, L. (2012). Phonological, visual, and semantic coding strategies and children's short-term picture memory span. *The Quarterly Journal of Experimental Psychology*, *65*(10), 2033–2053. https://doi.org/10.1080/17470218.2012.672997

Jarrold, C., & Citroën, R. (2013). Reevaluating key evidence for the development of rehearsal: Phonological similarity effects in children are subject to proportional scaling artifacts. *Developmental Psychology*, *49*(5), 837–847. https://doi.org/10.1037/a0028771

Jeffreys, H. (1961). *Theory of probability.,(Oxford University Press: Oxford, UK)*.

Keeney, T. J., Cannizzo, S. R., & Flavell, J. H. (1967). Spontaneous and Induced Verbal Rehearsal in a Recall Task. *Child Development*, *38*(4), 953–966. JSTOR. https://doi.org/10.2307/1127095

Logie, R. H. (2011). The functional organization and capacity limits of working memory. *Current Directions in Psychological Science*, *20*(4), 240–245.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281. https://doi.org/10.1038/36846

Morey, C. C., Hadley, L. V., Buttelmann, F., Könen, T., Meaney, J.-A., Auyeung, B., Karbach, J., & Chevalier, N. (2018). The effects of verbal and spatial memory load on children's processing speed: Development of WM load effects. *Annals of the New York Academy of Sciences*, *1424*(1), 161–174. https://doi.org/10.1111/nyas.13653

Morey, C. C., Mareva, S., Lelonkiewicz, J. R., & Chevalier, N. (2018). Gaze-based rehearsal in children under 7: A developmental investigation of eye movements during a serial spatial memory task. *Developmental Science*, *21*(3), e12559. https://doi.org/10.1111/desc.12559

Morey, R. D., & Rouder, J. N. (2015). *BayesFactor: Computation of Bayes Factors for Common Designs.* (R package version 0.9.12-2) [Computer software].

Overkott, C. S. R., Morey, C. C., & Souza, A. S. (2022, August 13). The developing impact of verbal labels on visual memories in children. https://doi.org/10.17605/OSF.IO/3WESD

Overkott, C., & Souza, A. S. (2021a). The Fate of Labeled and Non-Labeled Visual Features in Working Memory. *Submitted Manuscript*.

Overkott, C., & Souza, A. S. (2021b). Verbal descriptions improve visual working memory but have limited impact on visual long-term memory. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001084

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897X00366

Pratte, M. S., Park, Y. E., Rademaker, R. L., & Tong, F. (2017). Accounting for stimulus-specific variation in precision reveals a discrete capacity limit in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(1), 6–17. https://doi.org/10.1037/xhp0000302

R Development Core Team. (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/.

Sarigiannidis, I., Crickmore, G., & Astle, D. E. (2016). Developmental and individual differences in the precision of visuospatial memory. *Cognitive Development*, *39*, 1–12. https://doi.org/10.1016/j.cogdev.2016.02.004

Sense, F., Morey, C. C., Prince, M., Heathcote, A., & Morey, R. D. (2017). Opportunity for verbalization does not improve visual change detection performance: A state-trace analysis. *Behavior Research Methods*, *49*(3), 853–862. https://doi.org/10.3758/s13428-016-0741-1

Shimi, A., Kuo, B.-C., Astle, D. E., Nobre, A. C., & Scerif, G. (2014). Age group and individual differences in attentional orienting dissociate neural mechanisms of encoding and maintenance in visual STM. *Journal of Cognitive Neuroscience*, *26*(4), 864–877.

Shimi, A., & Scerif, G. (2015). The interplay of spatial attentional biases and mental codes in VSTM: Developmentally informed hypotheses. *Developmental Psychology*, *51*(6), 731–743. https://doi.org/10.1037/a0039057

Shimi, A., & Scerif, G. (2017). Towards an integrative model of visual short-term memory maintenance: Evidence from the effects of attentional control, load, decay, and their interactions in childhood. *Cognition*, *169*, 61–83.

Shimi, A., & Scerif, G. (2021). The influence of attentional biases on multiple working memory precision parameters for children and adults. *Developmental Science*.

Simmering, V. R. (2012). The development of visual working memory capacity during early childhood. *Journal of Experimental Child Psychology*, *111*(4), 695–707.

Simons, D. J., & Levin, D. T. (1997). Change blindness. *Trends in Cognitive Sciences*, *1*(7), 261–267.

Souza, A. S., Overkott, C., & Matyja, M. (2020). Categorical distinctiveness constrains the labeling benefit in visual working memory. *Submitted Manuscript*.

Souza, A. S., Overkott, C., & Matyja, M. (2021). Categorical distinctiveness constrains the labeling benefit in visual working memory. *Journal of Memory and Language*, *119*, 104242. https://doi.org/10.1016/j.jml.2021.104242

Souza, A. S., & Skóra, Z. (2017). The interplay of language and visual perception in working memory. *Cognition*, *166*, 277–297. https://doi.org/10.1016/j.cognition.2017.05.038

Sutterer, D. W., & Awh, E. (2016). Retrieval practice enhances the accessibility but not the quality of memory. *Psychonomic Bulletin & Review*, *23*(3), 831–841. https://doi.org/10.3758/s13423-015-0937-x

Swanson, H. L. (2017). Verbal and visual-spatial working memory: What develops over a life span? *Developmental Psychology*, *53*(5), 971.

Vales, C., & Smith, L. B. (2015). Words, shape, visual search and visual working memory in 3-year-old children. *Developmental Science*, *18*(1), 65–79. https://doi.org/10.1111/desc.12179

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 t Tests. *Perspectives on Psychological Science*, *6*(3), 291–298.