Registered Replication Report

Advances in Methods and Practices in Psychological Science
April-June 2021, Vol. 4, No. 2, pp. 1–20

DOI: 10.1177/25152459211018187
www.psychologicalscience.org/AMPPS

Multilab Direct Replication of Flavell, Beach, and Chinsky (1966): Spontaneous Verbal Rehearsal in a Memory Task as a Function of Age

Emily M. Elliott[1]*, Candice C. Morey[2]*, Angela M. AuBuchon[3]*, Nelson Cowan[4]*, Chris Jarrold[5]*, Eryn J. Adams[4], Meg Attwood[5], Büşra Bayram[6], Stefen Beeler-Duden[7], Taran Y. Blakstvedt[8], Gerhard Büttner[9,10], Thomas Castelain[11], Shari Cave[12], Davide Crepaldi[13], Eivor Fredriksen[8], Bret A. Glass[4], Andrew J. Graves[7], Dominic Guitard[4], Stefanie Hoehl[14], Alexis Hosch[15], Stéphanie Jeanneret[16], Tanya N. Joseph[2], Chris Koch[17], Jaroslaw R. Lelonkiewicz[13], Gary Lupyan[15], Amalia McDonald[7], Grace Meissner[3], Whitney Mendenhall[17], David Moreau[12], Thomas Ostermann[18], Asil Ali Özdoğru[6], Francesca Padovani[13], Sebastian Poloczek[9,10], Jan Phillip Röer[18], Christina C. Schonberg[15], Christian K. Tamnes[8], Martin J. Tomasik[18], Beatrice Valentini[16], Evie Vergauwe[16], Haley A. Vlach[15], and Martin Voracek[14]

*Lead authors

[1]Department of Psychology, Louisiana State University, Baton Rouge, Louisiana, USA; [2]School of Psychology, Cardiff University, Cardiff, Wales, UK; [3]Boys Town National Research Hospital, Omaha, Nebraska, USA; [4]Department of Psychological Sciences, University of Missouri, Columbia, Missouri, USA; [5]School of Psychological Science, University of Bristol, Bristol, UK; [6]Department of Psychology, Üsküdar University, Istanbul, Turkey; [7]Department of Psychology, University of Virginia, Charlottesville, Virginia, USA; [8]Department of Psychology, University of Oslo, Oslo, Norway; [9]Department of Educational Psychology, Goethe University Frankfurt Institute of Psychology, Frankfurt, Germany; [10]Centre for Individual Development and Adaptive Education of Children at Risk (IDeA), Frankfurt, Germany; [11]Instituto de Investigaciones Psicológicas (IIP), University of Costa Rica, San José, Costa Rica; [12]School of Psychology, University of Auckland, Auckland, New Zealand; [13]Neuroscience, Scuola Internazionale Superiore di Studi Avanzati (SISSA), Trieste, Italy; [14]Department of Psychology, University of Vienna, Vienna, Austria; [15]Department of Educational Psychology, University of Wisconsin, Madison, Wisconsin, USA; [16]Faculty of Psychology and Educational Sciences, University of Geneva, Geneva, Switzerland; [17]Psychology Department, George Fox University, Newberg, Oregon, USA; and [18]Faculty of Health, School of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

Abstract

Work by Flavell, Beach, and Chinsky indicated a change in the spontaneous production of overt verbalization behaviors when comparing young children (age 5) with older children (age 10). Despite the critical role that this evidence of a change in verbalization behaviors plays in modern theories of cognitive development and working memory, there has been only one other published near replication of this work. In this Registered Replication Report, we relied on researchers from 17 labs who contributed their results to a larger and more comprehensive sample of children. We assessed memory performance and the presence or absence of verbalization behaviors of young children at different ages and determined that the original pattern of findings was largely upheld: Older children were more likely to verbalize, and their memory spans improved. We confirmed that 5- and 6-year-old children who verbalized recalled

Corresponding Author:
Emily M. Elliott, Department of Psychology, Louisiana State University
E-mail: eelliott@lsu.edu

more than children who did not verbalize. However, unlike Flavell et al., substantial proportions of our 5- and 6-year-old samples overtly verbalized at least sometimes during the picture memory task. In addition, continuous increase in overt verbalization from 7 to 10 years old was not consistently evident in our samples. These robust findings should be weighed when considering theories of cognitive development, particularly theories concerning when verbal rehearsal emerges and relations between speech and memory.

Keywords
development, rehearsal, verbalization, memory, short-term memory, working memory, Registered Replication Report, open data, open materials, preregistered

Received 5/15/17; Revision accepted 3/25/21

Speech is intertwined with thinking to such a degree that imagining thinking without verbalization is difficult. Self-directed speech may be used to support cognition. Early theories of cognitive development stipulated that such goal-directed speech emerges as overt motor productions, which become internalized as children mature (Vygotsky, 1962). Flavell et al. (1966) documented the prevalence of young children's overt speech-related motor productions during a verbal short-term memory task. Their article has become a seminal work supporting theories of verbal memory and cognitive development more generally. Their clear and compelling results provided a rough timeline of the developmental trajectory of applying speech to memorizing: Children 5 to 6 years old almost never verbalized at any point during the memory task, but verbalization instances increased in 7- to 8-year-old and 10- to 11-year-old children. Flavell et al. interpreted this pattern as confirming the production deficiency hypothesis, meaning that before age 7, children usually do not even produce the helpful verbalizations that might mediate improved memory performance. These striking findings have constrained theories about how children approach memorizing and why memory spans for verbalizable information increase throughout childhood.

Since the publication of these seminal findings by Flavell et al. (1966), a similar developmental trajectory has been proposed for the use of self-directed speech in other cognitive domains. Using speech overtly to support problem-solving is believed to occur before the development of covert, inner speech (Alderson-Day & Fernyhough, 2015): Evidence suggests that by age 7, children use overt, private speech (i.e., not directed toward another individual) in service of task goals, including task switching (Kray et al., 2008) and problem-solving (Al-Namlah et al., 2006). Eventually, evidence for the use of speech in service of thought transitions from recordable private speech to other physical actions that hint at covert speech (e.g., lip movements) or indirect experimental effects that imply verbal recoding, such as phonological similarity effects (Henry et al., 2012).

Flavell et al.'s (1966) early and striking evidence of children's shift toward verbalizing memoranda around the age of 7 years has retained its influence largely because their study was conducted with impressive conscientiousness. Flavell et al. asked children to remember sequences of pictures representing common items. They took great care not to encourage the children to name the pictures during the trials. The experimenter indicated the picture sequence by pointing to the pictures, and during a 15-s retention interval, the child's eyes were obscured by a helmet with its visor down so that the experimenters could watch the child's face closely without risking nonverbal communication or awkwardness. With the visors raised again, the children responded by pointing to the pictures in serial order so that speech was never necessary during any part of the procedure. However, speech was also not restricted or explicitly discouraged. The main dependent variable of interest was whether the child spontaneously named the pictured items at any point during the trial. Rather than depend solely on overt speech, Flavell et al. additionally arranged for an experimenter practiced in lip-reading to watch the children during the task and record lip movements that were consistent with covert rehearsal of the pictures' names. Their careful design and detailed report inspired confidence that the researchers sincerely attempted to detect evidence of overt verbalization.

These unambiguous findings have spurred theory about how memory develops across childhood. Some researchers believe that the emergence of verbal rehearsal use around age 7 suggests that the working memory component specializing in verbal rehearsal, the phonological loop, typically matures around age 7 (e.g., Gathercole, 1998; Hulme & Tordoff, 1989; Kail & Park, 1994). Evidence about the emergence of phonological similarity effects, which could occur only when remembering picture stimuli if participants opted to verbally recode the images, is often seen as being consistent with this assumption (e.g., Conrad, 1971; Henry et al., 2012). However, it remains unclear whether phonological similarity effects in children younger than age 7 are small or nonexistent (Jarrold & Citroen, 2013; Jarrold et al., 2015). Assuming a qualitative shift to verbalization behavior due to the emergence of a phonological loop depends on demonstrating that phonological similarity effects are

truly absent before some shifting point. Detecting differences in the magnitude of phonological similarity effects is hindered by proportional scaling artifacts. Jarrold and Citroen (2013) have demonstrated that the size of these effects is proportional to the length of the memory list to be recalled. Because the youngest children recall the shortest lists, they would be expected to show the smallest effects, but proportionally, their small effects are equivalent to those of older children. Although these demonstrations stop short of showing that children younger than 7 actually show comparable effects of internal verbalization to children older than 7, they also cast doubt on decades of research that appeared to show robust differences in these effects with increasing age.

Given the concerns about what can be inferred from the development of word length and phonological similarity effects and about the status of the hypothesized correlation between speech rates and memory span in young children (Jarrold & Hall, 2013), the findings of Flavell et al. (1966) arguably constitute the strongest evidence for the notion that verbal rehearsal emerges around age 7. However, this finding was produced, to our knowledge, only twice: by the original Flavell et al. study and a subsequent extension by Keeney et al. (1967). Although these two studies were exquisitely well designed and informative, they were based on samples of 60 children from the greater Rochester, New York, region and 89 children from the St. Paul, Minnesota, region, respectively, during roughly the same period in time. Therefore, in the current report, we examined the evidence that holds a key place in theoretical memory development. Specifically, in this contemporary replication, we examined Flavell et al.'s findings in geographically diverse international samples to determine whether the original findings are generalizable to modern data-collection techniques (e.g., computerized task presentation). Furthermore, this many-labs direct replication provided (a) greater detail about when any shift toward verbalization during memory tasks likely occurs, (b) further evidence regarding recent questions raised about children's rehearsal behaviors before the shift, and (c) additional generalization of the findings to a larger and more diverse sample, which will reveal how stable the verbalization onset age is with respect to regional differences in the ages that children begin formal schooling.

The original study of Flavell et al. (1966) was a model of thorough, conscientious research design, full of detail that demonstrated how carefully the researchers thought through the implications of their testing procedure. The results inspired numerous follow-up studies (i.e., as of December 2018, Google Scholar listed 1,088 citations; Web of Science listed 518), but no one precisely replicated the essential components of the original study's methodology. Flavell et al. provided detailed tabular data in their report itself. Unfortunately, the tabular data available were insufficient for precise estimates of when children begin spontaneously verbalizing or for quantifying the large individual differences one would reasonably expect in this estimate. Furthermore, these tabular data do not allow for systematically linking emergence of speech behaviors with improved recall. Flavell et al. attempted to test this link but were forced to test it on only a subset of their data, divided via an arbitrary post hoc procedure that was explicitly motivated by the need to divide a small sample of 7- to 8-year-olds into roughly equal groups of children who produced speech and children who did not. We did not think these limitations resulted from Flavell et al.'s research design. Rather, much larger and more diverse samples would have been needed to draw general conclusions about the timing of any such shift toward verbalization, how much variability could be reasonably placed around it, and whether it would directly improve recall.

Furthermore, to the extent that participation in formal education influences the age at which children begin to apply verbal strategies for remembering, it is crucial to gain a broader perspective on direct verbalization evidence. The seminal work of Flavell and colleagues (1966) was conducted in the mid-1960s in Rochester, New York. Since that time, many studies on the development of short-term memory and working memory in children have been conducted, but regional differences in approaches to education have previously not been considered with regard to whether and when children engage verbalization in the service of memory. For example, the typical age at which children begin formal schooling has changed in the United States since Flavell et al.'s study and still differs from when children elsewhere begin formal schooling. Prekindergarten programs in the United States began to appear in the mid-1960s but did not become widespread until the late 1970s (Administration for Children and Families, 2019). By 2014, only 43% of U.S. 3-year-olds and 66% of U.S. 4-year-olds were enrolled in prekindergarten (U.S. Department of Education and National Center for Education Statistics, 2016). In contrast, in the UK, these percentages were strikingly higher, with 96.3% of 3- and 4-year-old children in preschool programs (Department for Education, 2017). Differences in formal education practices may lead to differences in the onset of proactive verbalization in children, which muddies conclusions one can draw from individual local studies. Collecting data about verbalization behaviors in large samples in a variety of locations affords the opportunity to discern how much regional variability there is in both verbalization behaviors and short-term memory measures and presents the chance to observe whether these behaviors are linked to regional differences in schooling.

We followed Flavell et al.'s (1966) protocol, including the use of a prop to obstruct vision during delay periods, with a few modern updates aimed at providing better

control of timing and randomization of stimuli, which was essential for controlling the administration of the tasks across multiple sites. However, we remained as true to the original procedure as possible. For example, in providing instructions to the children regarding their response to the stimuli, we used similar language (e.g., p. 287 of the original manuscript states, "First I point to this one, and then I point, etc."; for specific details, see the protocol on our OSF page, https://osf.io/pn4rk/). In addition, we increased the precision in ages in the potential inflection region. Flavell et al. tested 5-, 7-, and 10-year-old children who were kindergartners and second and fifth graders in the U.S. school system, respectively. We increased precision by adding 6-year-olds (i.e., first graders in the United States). We also presented stimuli and recorded responses via computer using lab .js, a browser-based tool for designing and sharing openly available experimental tasks (Henninger et al., 2019), to minimize differences in administration that could occur across sessions and labs. We also gathered information about educational experiences, including enrollment in preschool programs (also known as prekindergarten) and the ages at which children began formal schooling.

Method in the Original Study

Participants

Sixty children were tested, 20 each from kindergarten (mean age = 69 months), second grade (mean age = 93 months), and fifth grade (mean age = 129 months). There were 10 boys and 10 girls tested in each age group.

Materials and procedure

Participants began with two practice trials of two items each. They were shown two pictures, and the experimenter pointed to them in a specific order "at the rate of one point per 2 seconds" (Flavell et al., 1966, p. 289). There were two conditions: immediate recall and delayed recall (in which recall was delayed by 15 s after presentation of the final list item), and the children always indicated their response by pointing. The order of presentation of the immediate and delayed recall conditions was counterbalanced such that half of the children received immediate recall first, and half received delayed recall first. The picture naming tasks were always last. The testing lasted approximately 20 to 25 min. List lengths varied by age group, with the 5- and 7-year-old children receiving one list per delay condition at each length two to four and the 10-year-old children receiving one list at each length three to five.

The participant was semiblindfolded during the 15-s delay period using a "space helmet." The procedure included time to habituate to the feeling of wearing the space helmet and to understand the verbal commands of "Visor Up!" and "Visor Down!" needed to ensure that the participant could see the stimulus presentation but was undisturbed during the retention interval.

Then, after the recall portion, participants were asked strategy questions:

> When he pointed to the pictures, you knew you were supposed to try to remember them, so you could point to the same ones afterward. Right? What did you do to remember them? I mean, how did you go about trying to keep them straight in your head? (Flavell et al., 1966, p. 289)

These questions were meant to give children every possible chance of somehow indicating that they attempted to remember the pictures by verbally rehearsing their names.

Next, participants were asked to review the set of pictures and to provide verbal labels for each of the items used. The experimenter pointed to each picture, and the child was asked to label each one. It was simply naming one item at a time. Following this, participants performed another version of the delayed recall task again, referred to as "Point and Name." It followed the same procedure as the delayed recall task except that the participants overtly named the pictures during presentation and recall.

One of the experimenters was trained to observe the children and to look for mouth movements consistent with saying the names of the pictures in the set. This experimenter was not aware of the contents of the trial, nor could he see the pictures on each trial. The stimuli chosen were set to elicit "large and conspicuous mouth movements" that "were distinctive and discriminable," such as "pipe" and "flag" (Flavell et al., 1966, pp. 286–287). The complete set of stimuli included drawings of an apple, comb, American flag, two yellow flowers on a stem, a moon, an owl, and a pipe.

Method

We developed a modernized version of Flavell et al.'s (1966) experimental protocol to better facilitate organizing data collection across multiple sites while preserving the key components of the original work, such as using a pointing response, obscuring children's vision during retention to decrease the likelihood that they discovered their lip movements were being monitored, and ensuring that the researcher scoring the lip movements was blind to the stimuli presented. The stimuli were randomly selected without replacement at run time per participant and presented via computer rather than manually so that their timings were identical regardless of experimenter. During presentation, the set of seven pictures were

displayed in a horizontal row across the computer screen. The pictures to be remembered were highlighted for 2 s each, and then the set of pictures was randomized again before recall. The experimenter followed the participant's gestures with the mouse to record which picture the child pointed toward in real time. Minor changes were made to the pictorial stimuli (described below) to ensure that the stimuli in the set would be familiar to contemporary children as young as 5 years old. We included 6-year-olds in addition to 5-, 7-, and 10-year-olds to get a more precise estimate of when a shift toward verbalization might occur. Finally, there was no performance-based stopping rule. Instead, in each list-memory phase, two lists were presented at each list length used (two, three, four, and five items), as opposed to just one list per length in the original procedure. These changes lengthened the session somewhat, but importantly, they ensured that each child completed the same number of trials and therefore had the same number of opportunities to demonstrate verbalization behavior.

After the immediate and delayed recall tasks, the experimental program prompted the children to describe their approach to remembering the pictures in an open-ended response that was entered by the experimenter. This followed the original procedure of Flavell et al. (1966) and used the same language as Flavell et al. reported in their article. We added more specific follow-up prompts that could be endorsed with a yes or no response if there was additional discussion by the child about the strategies that they used (e.g., I said the pictures to myself, one at a time), but these prompts were not required.

Participants

We encouraged participating labs to recruit 80 children in total: 20 each of 5-, 6-, 7-, and 10-year-olds. In pilot testing, the experimental session lasted 25 to 30 min. To avoid bias and p-hacking, we advised every contributing laboratory to aim to collect 20 participants in each of the four age groups so that each lab-level study may reasonably replicate Flavell et al. (1966), but we allowed labs to collect larger samples whenever feasible, provided that the lab documented that any classical inferential analyses performed did not occur before all data collection ended. No instances of data-peeking were reported to the analyst when recruitment ended. Minimally, each lab committed to recruiting eight children per age group. Regardless of sample size, labs were asked to recruit the same numbers of boys and girls in each age group. Because this combined effort from multiple labs afforded a meta-analytic estimate as well as individual replications, we did not consider it necessary for each individual attempt to have higher statistical power than Flavell et al. had.

Data from labs contributing at least eight children in the relevant age groups for an analysis were included in lab-level analyses; all data were included in analyses in which identifying a per-lab effect was not the aim. We recruited typically developing children. Therefore, exclusionary criteria included having a developmental delay or neurocognitive disorder, hearing loss, or other significant health problem that could cause a discrepancy between mental age and chronological age or difficulty with pointing (e.g., an arm in a cast or sling).

Stimuli and software

Given the international effort to replicate the original study, two changes were made to the original pictorial stimuli. The American flag was replaced by a generic blue flag, and the pipe was replaced by an image of a pencil. The pictorial stimuli were taken from the Bank of Standardized Stimuli (Brodeur et al., 2010, 2014). For the specific file names of the images we used, see our online supplementary materials (https://osf.io/gwxq7/wiki/home/).

The program was created using lab.js, which is graphical interface for creating Javascript experiments (Henninger et al., 2019).

Local procedures

Every participating laboratory followed a common protocol based on Flavell et al.'s (1966) procedure that incorporated the updates we described above. Our lab .js program ensured that stimulus presentation and response recording were standardized across labs. There were nonetheless minor local differences between the settings in which the data were collected, which are described in more detail on the OSF project page (https://osf.io/pn4rk/). Ethical regulations governing data collection per locale varied somewhat (e.g., as to whether video recordings could be made); further details are provided in the individual procedure descriptions.

Results

Summary of deviations from the preregistered plan

The analyses we report below deviate from our Stage 1 plan in the following ways. First, we provide per-lab participant exclusions as a summary in text and details on the individual lab procedures in our online supplement rather than in our demographic table (Table 1). We chose to alter this to keep Table 1's size manageable.

Second, we added calculations of interrater reliability of the children's qualitative reports of the strategies they used to complete the task. This step was mentioned in our preregistered data-processing script but not in our Stage 1 manuscript. Our most substantial alteration involved data preparation for the analyses investigating whether spans and speech behavior during the delay period increased under the point-and-name instructions. Before data collection, we were concerned that the researcher coding speech behaviors during the point-and-name task would have been biased by hearing the child articulate the picture names during encoding; this would mean that unlike in the delayed recall task, the coder would always have had information about what the correct list sequence was, which may have influenced the coder's judgments. We therefore thought to restrict comparisons of the point-and-name and delayed recall data to sessions that were video-coded so that coders could observe and rate lip movements without sound. However, we ultimately chose not to request this extra coding on the subset of our data that afforded it. Researchers reported that overt speech was far more prevalent than speechless lip movements, which means that coders would not usually have resorted to lip movements when coding during either kind of trial. Interrater agreement (as available) was usually quite high, which suggests that coders did not have trouble performing the mandated task of identifying whether at least one speech instance was observed per period. These factors led us to conclude that recoding videoed point-and-name trials would not be worthwhile. In addition, because our procedure emphasized regularity across participants (including computer-controlled timing of the stimulus presentation), many children were observed to struggle with naming the pictures aloud on tempo; some failed to do it without reminders. In these frequent cases, the coder could not have been biased by hearing all of the names in correct order during presentation. Given this unanticipated difficulty and the limitations it places on interpreting this analysis, we decided to simply report the planned analysis comparing point-and-name and delayed recall without further restriction. For a complete guide to updates we made to our Stage 1 plan in the Stage 2 manuscript, see our OSF page (https://osf.io/vy39r/).

Lab demographics

The samples collected from each participating laboratory are summarized in Table 1.

Data processing

Demographic exclusions. In the case that individual laboratories ran participants who did not meet the study's explicit inclusion criteria (e.g., because the sibling of a recruited participant also wanted to take part), these participants were not considered part of the sample and were not entered into any of the analyses we report here.

Picture-naming exclusions. Flavell et al. (1966) began their analyses by discussing the findings from the picture-naming portion to be certain that any errors in recall were not due to a problem with production or the simple inability to name the stimuli. They described the picture-naming performance, documented that only a few of the participants in the youngest age group showed any trouble with naming the pictures, and continued with their analysis. On the basis of Flavell et al.'s report, we anticipated that only a few children would be unable to name the pictures. However, we judged that it would be most fair to exclude participants who did not demonstrate knowledge of the picture's names. Children who gave no response or "I don't know" to a picture prompt were excluded. We accepted "minor" labeling errors such as those Flavell et al. described as adequate because as long as the child had a reasonable label for the picture, verbalization could be detected (e.g., such as saying "brush" instead of "comb"). Each lab independently determined whether a participant's responses to the picture-naming prompts reflected adequate labels. Flavell et al. would have excluded two 5-year-olds under these criteria. Because we anticipated few exclusions and collected large samples overall, we did not require labs to replace participants excluded because of failures to name the pictures. Sixty-one children (0%–16% per sample) out of 977 (6%) failed to name at least one of the pictures adequately and were excluded from the analyses.

Coding of lip movements. At least one of the experimenters per lab was trained to recognize the lip movements associated with the object names of the seven picture stimuli (see our standardized training protocol, https://osf.io/36ayh/). Raters were instructed to indicate on each trial, separately during presentation, delay, and recall (as applicable per recall task), whether the child demonstrated verbalization behavior. The rater marked the event 1 if they observed indisputable evidence of stimulus labeling (e.g., hearing a specific word that matched one of the stimulus labels, lip-reading a stimulus word with certainty, or both). The rater assigned a 2 when they were reasonably certain that the child was labeling but did not identify a specific stimulus word through either hearing or lip-reading. Behaviors that justified a 2 included lip movements consistent with stimulus labeling or murmuring. Raters assigned a 3 when there was no discernible speech or lip movement or when any lip movement was definitely not speech related (e.g., pursing lips, clicking tongue, humming) or was irrelevant to the task (e.g., asking the

Table 1. Participant Demographics by Lab

				Age in months		Formal schooling	
University	Group	*N*	% Male	Mean age	*SD* (age)	Mean years in school	*SD* (school)
Boys Town National Research Hospital	5yo	20	50	65	3.70	1.50	0.89
	6yo	20	50	79	3.30	2.00	0.92
	7yo	20	50	91	3.60	2.90	0.85
	10yo	20	50	125	3.48	6.00	0.79
Cardiff University	5yo	19	53	66	2.61	0.89	0.58
	6yo	13	46	79	3.22	1.91	0.70
	7yo	17	53	88	2.83	2.93	0.46
	10yo	20	60	124	3.52	5.67	0.69
George Fox University	5yo	11	55	67	2.66	0.00	0.00
	6yo	7	43	79	3.16	0.71	0.49
	7yo	6	33	90	2.74	1.83	0.41
	10yo	8	50	126	3.52	5.00	0.00
Goethe University Frankfurt am Main	5yo	37	54	68	2.66	0.00	0.00
	6yo	49	45	77	3.67	0.00	0.00
	7yo	45	49	89	3.76	0.24	0.43
	10yo	27	56	124	2.91	3.37	0.49
Louisiana State University	5yo	9	56	66	3.87	0.71	0.95
	6yo	10	40	76	3.98	2.30	0.95
	7yo	12	33	92	3.90	3.33	1.15
	10yo	8	50	126	3.31	6.38	0.52
Scuola Internazionale Superiore di Studi Avanzati (SISSA)	5yo	8	50	66	4.03	1.12	0.64
	6yo	13	54	79	3.50	1.31	1.11
	7yo	22	41	89	3.72	2.24	0.70
	10yo	12	50	125	2.84	5.33	0.89
University of Auckland	5yo	18	56	64	3.37	0.00	0.00
	6yo	18	56	78	2.70	1.00	0.00
	7yo	20	50	91	2.39	2.00	0.00
	10yo	20	50	127	3.84	5.00	0.00
University of Bristol	5yo	3	100	64	4.58	1.00	0.00
	6yo	8	38	77	3.11	2.00	0.00
	7yo	9	44	90	3.24	3.00	0.00
	10yo	10	50	124	3.86	6.00	0.00
University of Costa Rica	5yo	18	50	65	3.34	0.78	0.43
	6yo	20	50	78	3.69	2.00	0.00
	7yo	20	50	90	3.43	3.00	0.00
	10yo	20	50	127	2.62	5.00	0.00
University of Geneva	5yo	10	60	66	2.77	1.00	0.47
	6yo	19	79	77	3.14	1.68	0.48
	7yo	13	54	88	3.06	2.69	0.63
	10yo	12	33	127	2.07	5.58	0.51
University of Missouri	5yo	14	57	65	3.56	1.17	1.11
	6yo	23	35	78	3.74	1.76	1.18
	7yo	17	41	88	3.33	2.73	1.28
	10yo	19	53	125	3.98	5.62	1.20
University of Oslo	5yo	14	50	64	3.79	NA	NA
	6yo	10	60	76	3.80	NA	NA
	7yo	15	47	90	3.76	NA	NA
	10yo	12	58	123	2.90	NA	NA
University of Vienna	5yo	8	50	66	2.73	0.00	0.00
	6yo	8	50	78	3.60	0.00	0.00

(continued)

Table 1. *(continued)*

University	Group	*N*	% Male	Age in months		Formal schooling	
				Mean age	*SD* (age)	Mean years in school	*SD* (school)
	7yo	9	44	91	3.47	0.56	0.53
	10yo	8	50	129	2.14	3.50	0.53
University of Virginia	5yo	9	44	66	4.11	NA	NA
	6yo	9	44	79	3.00	NA	NA
	7yo	10	50	90	3.68	NA	NA
	10yo	10	60	126	3.37	NA	NA
University of Wisconsin	5yo	8	62	65	3.12	1.43	0.79
	6yo	12	25	76	3.73	2.55	0.82
	7yo	19	58	88	3.61	3.39	0.98
	10yo	13	46	123	4.07	6.42	1.00
University of Witten/Herdecke	5yo	8	50	64	3.66	0.00	0.00
	6yo	8	50	81	3.34	0.00	0.00
	7yo	8	50	86	3.07	1.00	0.00
	10yo	8	50	125	3.74	4.00	0.00
Üsküdar University	5yo	6	33	65	3.93	0.67	0.82
	6yo	6	67	79	4.13	1.00	0.00
	7yo	7	57	90	3.74	2.14	0.38
	10yo	8	50	124	3.48	5.50	0.76

Note: We relied on local definitions of formal schooling. In some regions, play-based care for young children includes little or no formal instruction; formal instruction is widely acknowledged to begin with a specific program started by all children at a particular age. In other regions, provision is much more mixed, with some preschools administering prereading instruction within predomoinately play-based day care. Parents in these regions were likely to consider that their child's formal education had begun. Values in the mean years of formal schooling column depended on the parents' definition of the sort of instruction that would constitute formal schooling. yo = years old; NA = not applicable.

experimenter a question). Responses 1 and 2 were accepted as evidence of speech, as in Flavell et al. (1966).

Coding of strategy-free response. After performing the recall tasks, participants were asked about how they remembered the pictures. The first of these strategy questions was open-ended, and researchers typed the participant's response for later scoring. This open-ended question was followed up by yes-or-no questions that asked about specific strategies, including verbalization, more directly. When possible, two researchers reviewed the free responses and rated whether the responses indicated unambiguous reliance on covert verbalization. To be considered unambiguous, the child needed to indicate that the strategy used was phonological in nature by referencing, for example, their use of "words," "names," "sounds," "said," or specific labels (e.g., apple, moon). The protocol specifically contrasted these examples with potential responses of "I thought about them" and "I tried to remember them," neither of which were considered verbalization strategies.[1]

Reliability of categorically coded verbalization behaviors and strategy-free responses. When possible, we video-recorded sessions or involved multiple raters in the live session to allow for double-coding of lip movements and speech behaviors. This allowed us to calculate interrater reliability estimates for some samples to increase confidence in the results. However, not all labs were equipped to record sessions or duplicate testing personnel, and individual children did not always consent to be recorded. Nonetheless, the subset of the data that were coded by multiple raters (available for 12 of our 17 participating labs for verbalization ratings and 16 out of 17 for strategy-free responses) provides some estimate of rater reliability. All recordings of children remained local and were not shared to protect the privacy of the children who participated. We computed Cohen's κ scores to assess interrater reliability. Because we do not know how reliable Flavell et al.'s (1966) ratings were, we proceeded with our analyses regardless of the κ value obtained. Procedures for resolving interrater disputes were decided locally per laboratory and may be found in the Local Procedures section. The κ values on both verbalization behaviors (range = 0.70–1, *Mdn* = 0.95) and strategy-free responses (range = 0.33–1, average = 0.92) suggest that individual rater judgments of speech behaviors were quite objectively determinable.

Classification and description of verbalization behaviors

In their Table 1, Flavell et al. (1966) indicated whether participants in each age group spoke zero times, one to two times, or three or more times for both immediate

Table 2. Number of Children Showing Evidence of Verbalization by Age Group

Group	Never	Sometimes	Usually	*N*
5yo	53	80	87	220
6yo	28	84	141	253
7yo	19	73	177	269
10yo	14	40	181	235

Note: yo = years old.

and delayed recall trials combined. We modified this to maintain consistency with the original report because we increased the number of trials each child experienced. We ran eight trials each in the immediate and delayed recall procedures (each child received two lists at list lengths two to five). With only three trials each in the immediate and delayed recall tasks, Flavell et al. categorically divided children into three categories according to their speech behavior: They evinced zero instances of verbalization, one to two instances of verbalization, or more than three instances of verbalization. To make a comparable table, we similarly categorized children as speaking never (on zero out of 16 trials, in any period of the trial), sometimes (on one to seven trials), or usually (eight or more trials). Flavell et al. counted both the "indisputable" (i.e., assigned rating 1) and "reasonably certain" (i.e., assigned rating 2) levels as indicating relevant speech production. We likewise used this lenient scoring but reproduced all contingent analyses using a stricter designation in which only the rating indicating indisputable evidence of verbalization was considered speech production. The stricter analysis was necessary for assessing how robust our conclusions were to these rating decisions. For results of the stricter analysis, see our online supplement (https://osf.io/gqym3/); outcomes were comparable with those reported here. For analyses in which we needed a binary representation of whether a child was a speech producer, we leniently considered participants who sometimes produced speech as producers.

The numbers of participants who verbalized in each age group are displayed in Table 2, which is similar to that of Table 1 from Flavell et al. (1966). To show how samples from individual labs looked, we present the individual lab findings as heat maps (Fig. 1; darker colors indicate larger proportions of instances per verbalization behavior category). These heat maps indicate how closely each replication matched Flavell et al.'s original data, in which verbalization transitioned from little or no speech in the youngest group to frequent speech in oldest group. Visual inspection of these heat maps verifies that Flavell et al.'s original observation of increasing verbalization with age was consistently observed across the 17 participating labs. However, note that we also observed much higher proportions of verbalization in 5-year-old children than Flavell et al. documented.

Sample-wide χ^2 analyses were conducted to replicate Flavell et al.'s (1966) findings. Like Flavell et al., our omnibus χ^2 test on these values revealed a significant shift across ages in speech behavior, $\chi^2 = 87.47$, $df = 6$, $p < .01$. Table 3 provides pair-wise χ^2 outcomes comparing the 5- and 7-year-olds, 5- and 10-year-olds, and 7- and 10-year-olds (the same ages of children Flavell et al. tested; we excluded 6-year-olds from these replication analyses). We used Bonferroni correction for multiple comparisons and considered statistically significant only *p* values less than .01. Accordingly, like Flavell et al., our omnibus analysis suggests shifts in verbalization behavior between 5- and 7-year-olds and 5- and 10-year-olds but narrowly misses the criteria for a shift between 7- and 10-year-olds. Bayes factor contingency tests (Morey et al., 2018) support the same inference: The Bayes factors in Table 3 similarly suggest clear evidence for differences between 5-year-olds and older children but neither confirm nor disconfirm shifts in verbalization between 7- and 10-year-olds.

For synthesizing effect sizes across labs, we fit random effects models using the *metafor* package (Viechtbauer, 2010). We considered a child a speaker if the child ever seemed to verbalize during the memory task (i.e., considering sometimes and usually participants as speakers and never participants as nonspeakers). We separately compared 5- and 7-year-olds (see Fig. 2), then 7- and 10-year-olds (Fig. 3). The forest plots in Figures 2 and 3 depict the effect sizes per lab group with 95% confidence intervals, and the meta-analytic effect size is represented by the diamond at the bottom of the figures. Tests of effect size differences are given in Table 4; these are consistent with the assumption that the variability in effect sizes between labs was random. For 5- and 7-year-old children, most labs observed increases in speech behavior consistent with those reported by Flavell et al. (1966): Here, the confidence intervals around the meta-analytic effect size did not include zero. The evidence was much less clear when we considered the transition from 7 to 10 years old: Some individual labs observed increases in proportions of participants verbalizing, but many also did not, and intervals around the meta-analytic effect size included zero.

Verbalization behaviors by subtask periods

We also divided each task into intervals corresponding to presentation, delay, and recall and recorded verbalization behaviors during these three intervals. These data are presented in Table 5, which is similar to Table 2 from Flavell et al. (1966) except that proportions rather than

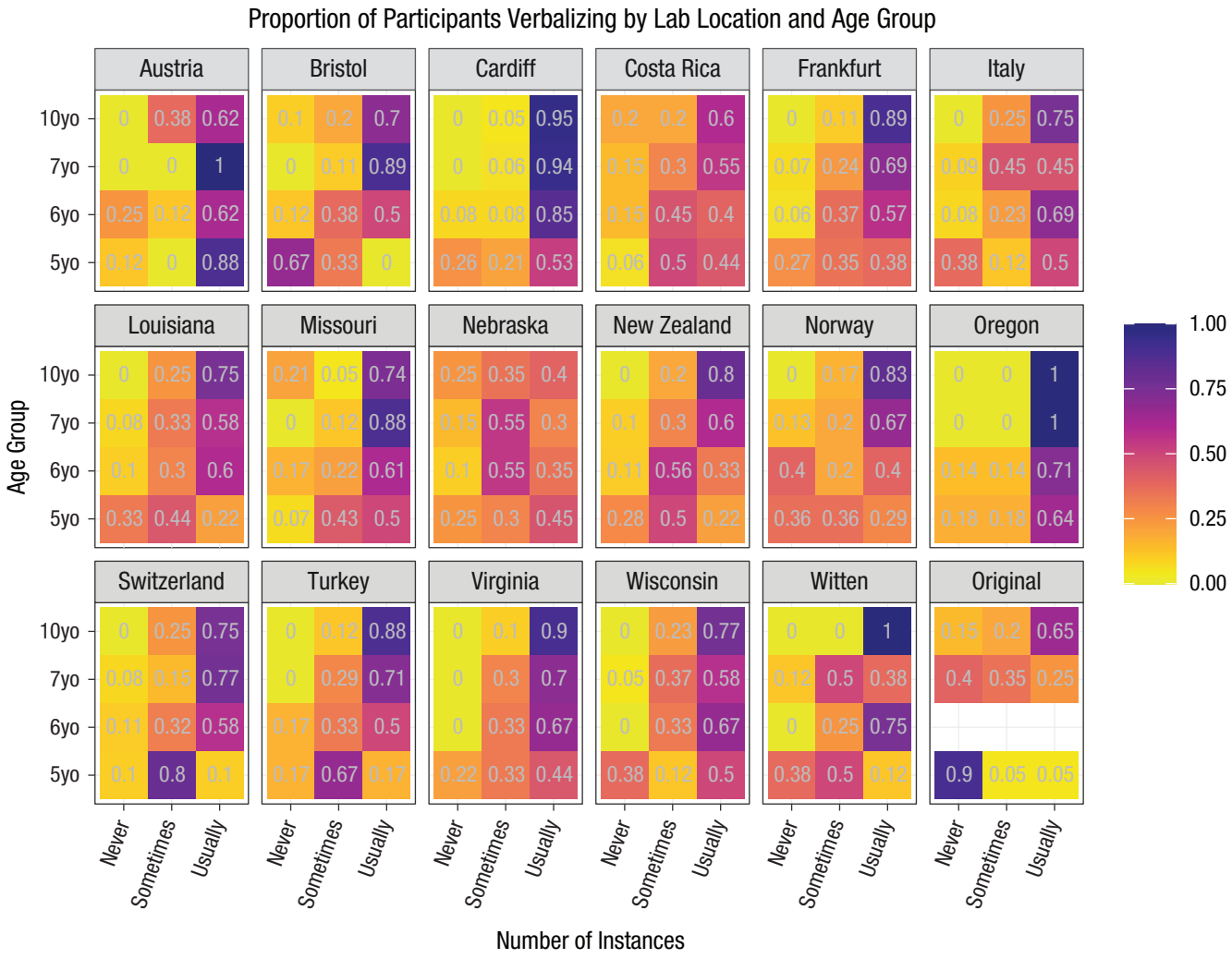

Fig. 1. Proportion of participants verbalizing by lab location and age group using the categories of never, sometimes, and usually to indicate the amount of verbalization. Lab sample sizes are given in Table 1.

raw numbers of participants are given. Flavell et al. made two points about these data, both of which were informally supported by our descriptive values. First, they noted that participants did not seem to be more likely to speak during the delay period even though that is when they had the longest opportunity to speak. We also did not observe numerically more speech observations during delay than during presentation or recall periods. Second, Flavell et al. observed that participants appeared more likely to speak during the point-and-name delay period than during the delay period of the delayed recall task. Flavell et al. did not have large enough samples to do persuasive inference on these values. Using our entire sample, we carried out a trial-level logistic regression on speech behavior per task, age group, and trial period with participant specified as a random effect nested within the random effect of lab. We ran this more complicated analysis to accommodate the nested structure (e.g., participants were part of one of 17 unique samples) of our data, but our intent with this analysis was merely to confirm whether Flavell et al.'s descriptive claims were inferentially supported. For simplicity, we therefore report the summary statistics given by the *anova* function from the R package *lmerTest* (Kuznetsova et al., 2020), as advised by Luke (2017), applied to *lme4* output (Bates et al., 2020). The omnibus model revealed significant effects (ps < .05) of age group (F = 32.24, df = 3), task (F = 51.35, df = 2), and period

Table 3. Pairwise χ^2 Tests Comparing Number of Speakers per Age Group

Ages	χ^2	*df*	*p*	Bayes factor
5- and 7-year-olds	42.58	2	.00	3.65×10^7
5- and 10-year-olds	68.59	2	.00	3.076×10^{13}
7- and 10-year-olds	8.18	2	.02	0.6097

Note: The *p* values are rounded to two decimal places.

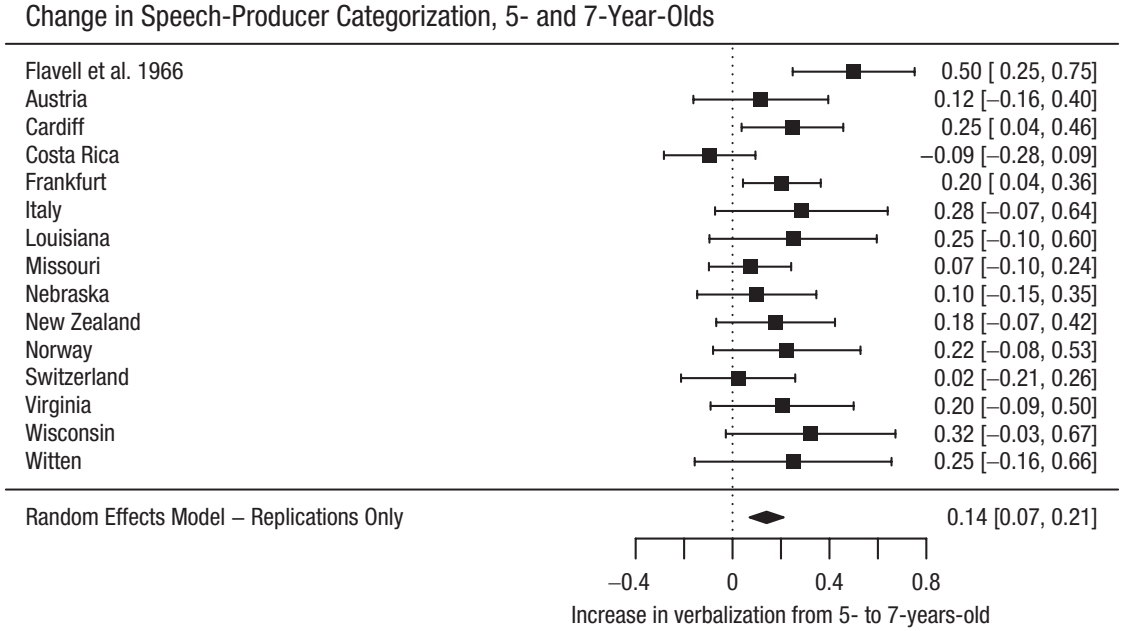

Fig. 2. Effect sizes per lab and meta-analytic effect size (with 95% confidence intervals) comparing 5- and 7-year-olds showing any as opposed to no speech behavior. Labs with samples fewer than eight in either age group were omitted from this analysis.

within task ($F = 1{,}069.12$, $df = 2$) and Age Group × Task ($F = 8.76$, $df = 6$) and Age Group × Period ($F = 19.39$, $df = 6$) interactions. Each factor discussed by Flavell et al. indeed appears inferentially supported, so we carried out further analysis needed to support these claims.

To confirm Flavell et al.'s (1966) finding that speech during the delay period was not more likely than during presentation or recall periods of the delayed recall test, we carried out a follow-up analysis focusing on the delayed recall task only. In our replication, speech was

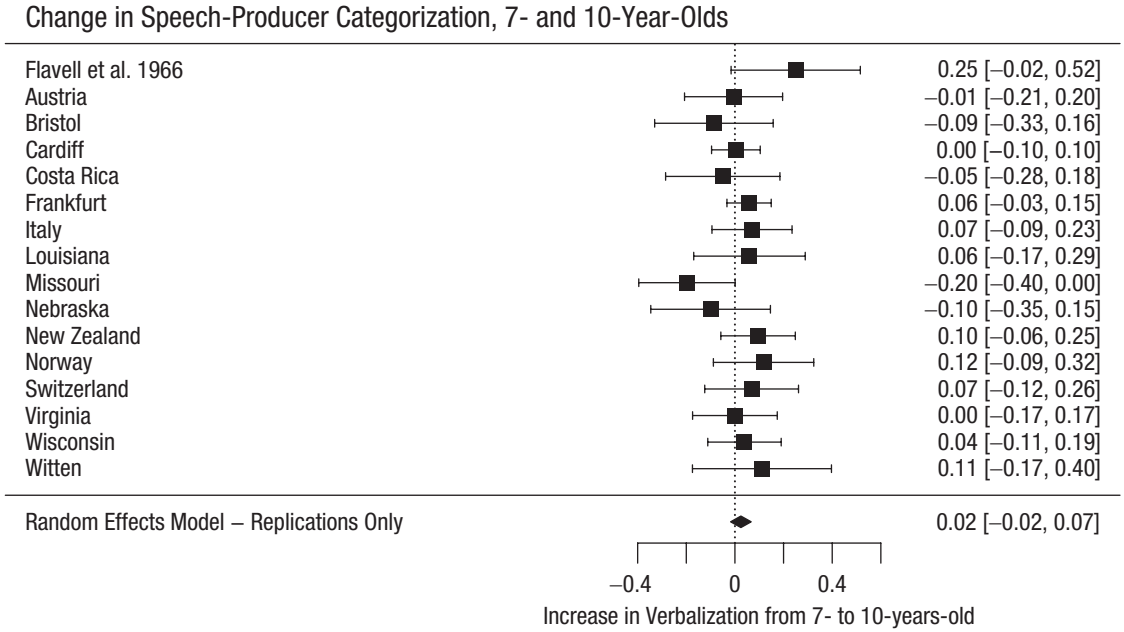

Fig. 3. Effect sizes per lab and meta-analytic effect size (with 95% confidence intervals) comparing 7- and 10-year-olds showing any as opposed to no speech behavior. Labs with samples fewer than 8 in either age group were omitted from this analysis.

Table 4. Meta-Analysis Heterogeneity Statistics

Analysis	τ^2	I^2	H^2	Q	*p*
Any vs. no speech, 5- and 7-year-olds	.002	12.936	1.149	12.110	.519
Any vs. no speech, 7- and 10-year-olds	.000	0.000	1.000	10.412	.732

observed less frequently during the delay period (M = 0.26, SD = 0.44) than during presentation (M = 0.34, SD = 0.47) or recall (M = 0.50, SD = 0.50; F = 759.32, df = 2),[2] as it seemed to be in Flavell et al.'s sample.

To confirm that participants were more likely to speak during the point-and-name delay period than during the delay period within the delayed recall task, we carried out an analysis focusing on the delay period only. Speech was indeed observed less frequently during the delay period of the delayed recall task than during the delay period of the point-and-name task (M = 0.37, SD = 0.48; F = 347.54, df = 1).

Self-report of internal verbalization

The original authors (Flavell et al., 1966) also evaluated the observations of verbalization relative to the strategies that participants reported during the "Inquiry" section of the task (Flavell et al., 1966, Table 3). They noted that behaviors were largely consistent across observational data and the self-reported answers to the strategy questions; however, 15 participants yielded inconsistent patterns. These were then divided into two sets: eight children who showed behavioral evidence of verbalization without the matching self-report and seven children who said they had verbalized but had not been noted by the experimenter as producing overt speech. If the criteria were to include both researcher observation of verbalization and self-report of verbalization when classifying children as producing speech, the number of 5-year-old children who produced speech was still very small (i.e., two) and was much larger in the 7- and 10-year-old children. We replicated Flavell et al.'s Table 3 across our entire sample (Table 6) and allowed for reconsideration of totals of producers and nonproducers when participants who self-reported internalized verbalizing were considered as producers along with those who were observed performing speech behaviors. The values in Table 6 are consistent with those reported by Flavell et al. in that very few 5-year-olds who did not evince speech during the session reported using verbalization to remember the pictures. This value was not very different in the older children (see Table 6), but we observed few participants older than the age of 5 out of a very large sample who never seemed to verbalize during the tasks. When we account for self-reports, a fair number of 5- and 6-year-olds still show no evidence of verbalizing, but vanishingly small numbers of 7- and 10-year-olds do not acknowledge verbalizing by report or observation.

Relationships between verbalizing and memory span

Figure 4 shows the average maximum span length recalled correctly per age group and lab; unsurprisingly,

Table 5. Proportions of Participants Spontaneously Speaking During Each Segment of Each Subtask by Age Group

Group	Recall task	Presentation-original	Delay-original	Recall-original	Presentation-replication	Delay-replication	Recall-replication
5yo	Delayed	0.00	0.10	0.05	0.43	0.39	0.51
	Immediate	0.05	NA	0.05	0.45	NA	0.54
	Point and name	NA	0.35	NA	NA	0.61	NA
6yo	Delayed	NA	NA	NA	0.62	0.56	0.70
	Immediate	NA	NA	NA	0.60	NA	0.67
	Point and name	NA	NA	NA	NA	0.72	NA
7yo	Delayed	0.35	0.35	0.35	0.71	0.69	0.83
	Immediate	0.20	NA	0.25	0.69	NA	0.82
	Point and name	NA	0.65	NA	NA	0.80	NA
10yo	Delayed	0.30	0.50	0.60	0.78	0.66	0.85
	Immediate	0.55	NA	0.65	0.72	NA	0.83
	Point and name	NA	0.80	NA	NA	0.83	NA

Note: The immediate recall task had no delay period, and verbalization behaviors were explicitly instructed (so were not spontaneous) during the presentation and recall periods of the point-and-name task. yo = years old; NA = not applicable.

Table 6. Number of Participants Reporting Internal Speech by Observed Speech Behavior and Age Group

Observed speech	Reported speech	5yo	6yo	7yo	10yo
No	No	39	15	6	2
No	Yes	7	10	10	12
Yes	No	111	88	70	21
Yes	Yes	40	115	157	174

Note: Some responses were missing. yo = years old.

this shows remarkably consistent increases in memory span with age across the participating laboratories. Flavell et al. (1966) were interested in whether children who were observed speaking remembered more than children who were not. They divided children according to the presence or absence of speech behaviors to test whether children who spoke also recalled longer lists. Because in their sample the youngest children so rarely verbalized and the oldest children always provided some evidence of verbalization, whether via overt speech behavior or self-report, Flavell et al. had to restrict analysis of verbalization behavior on memory span to only the intermediate group of children. To run a reasonable *t* test, they had to divide their 7-year-old sample in whatever manner they could justify that resulted in the creation of two roughly equal groups. They chose to divide the 7-year-old children into groups according to whether they verbalized during recall ($N = 11$) or not ($N = 9$) and found that children who verbalized recalled longer lists than children who did not. Note that this ad hoc definition of verbalization as overt speech or lip movements specifically observed during recall differed from how they categorized speech producers elsewhere in their original article. Verbalization during recall, as opposed to during presentation or during the delay period of delayed recall, also seems the least likely to reflect use of internal verbalization to assist with memory; verbalizing only at recall could merely reflect that

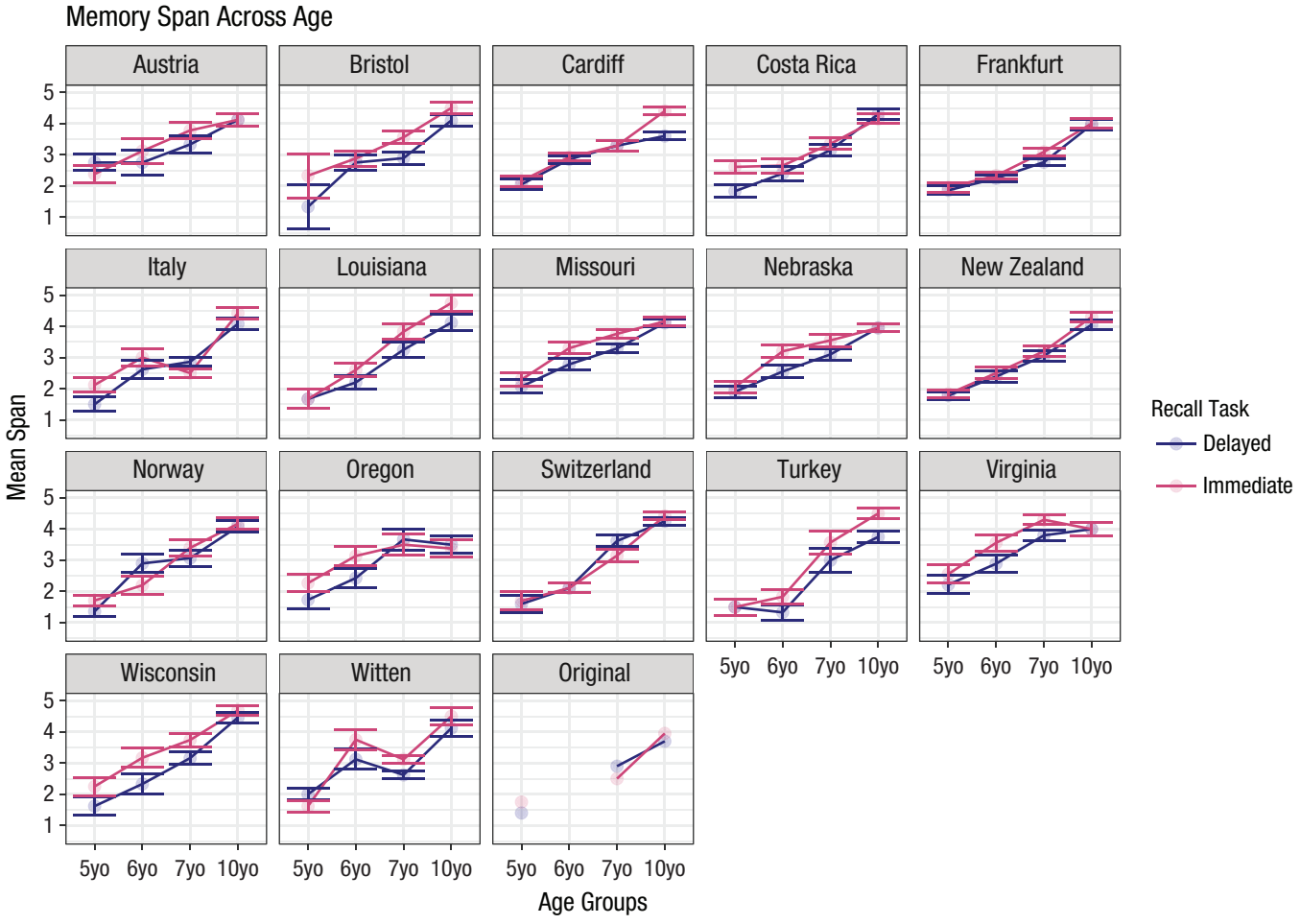

Fig. 4. Memory spans by age group and recall task per lab group. Error bars are within-participants standard errors of the mean (Morey, 2008). Children who never responded correctly were assigned a span score of 1, consistent with Flavell et al. (1966).

children realize that speaking is a convenient way to communicate with the researcher.

It would clearly be much better to define these groups a priori and consistently from one analysis to the next. Rather than replicate Flavell et al.'s (1966) *t* tests exactly, our preregistered analysis plan called for what we suppose Flavell et al. would have done had their sample sizes permitted. We analyzed span scores (defined as the maximum sequence length a child correctly recalled, as in Flavell et al.) using the lenient identification of the three-level verbalization frequency assignment from Table 2 (i.e., children falling into the sometimes or usually categories from Table 2 were speech producers, and children in the never category were nonproducers), age group was between-participants factors, and recall task (delayed recall, immediate recall) was a within-participants factor. We preregistered both classical and Bayes factor (BF) analyses of variance (ANOVAs; Morey et al., 2018) for this analysis because our large samples make observing a significant effect potentially trivial in terms of effect size. Although Flavell et al. ran their comparable analysis only on the 7-year-old sample, we planned to run this analysis on each age group for which we had samples sizes of at least 20 in both the speech producer and nonproducer categories after combining data from each participating laboratory to ensure that there were sufficiently large numbers of participants per cell for a meaningful outcome. We observed at least 20 speakers and nonspeakers in the 5- and 6-year-old age groups (see Table 6; < 20 of the 7- and 10-year-olds refrained from producing speech); therefore, we included only 5- and 6-year-olds in our analysis, which is consistent with our preregistered plan. The average data corresponding to this analysis are shown in Figure 5. According to the Bayesian ANOVA, the best model (BF = 2.1×10^{27}) included main effects of each factor.[3] Including the effect of age group was favored by a factor of 2.2×10^{9}. Including the effect of verbalization frequency was favored by a factor of 3.1×10^{11}. Including the effect of recall task was favored by a factor of 2,668. Excluding interaction terms was favored by at least a factor of 4. Excluding the crucial Age Group × Verbalization Frequency interaction (which would be necessary to argue that children younger than 6 do *not* benefit from speech) was favored by a factor of 8 (i.e., there was substantial positive evidence for the null hypothesis here, which Bayesian analyses can provide). Thus for both 5- and 6-year-olds, the qualitative increase in verbalization corresponded to increases in remembered items.

Consistent with their extended discussion of whether verbalization might enhance memory (Flavell et al., 1966, pp. 295–296) as well as their explicit hypotheses (Flavell et al., 1966, p. 290), Flavell et al. (1966) considered whether performance in the point-and-name condition might be viewed as an intervention that could boost recall performance or increase observations of verbalizing during the retention interval compared with delayed recall without explicitly instructed picture naming. Flavell et al. simply noted that participants consistently verbalized more during retention in the point-and-name condition without providing statistical inference. They also noted that because it was always performed last, any increase in point-and-name performance could be attributed to practice effects and supported this contention by showing that order of the two experimental tasks did influence the amount of material recalled. We did not consistently observe this order effect: It interacted with age. Specifically, an analysis of variance on maximum spans including age group and experimental task order (immediate recall first or delayed recall first) as factors uncovered significant main effects of both factors and a significant interaction, $F(3, 1946) = 3.70$, $p < .05$. Only for the 10-year-old group was a numerically higher average observed for the second task ($M = 4.21$, $SD = 0.79$) than the first task ($M = 4.06$, $SD = 0.90$); for all other ages, we observed the reverse, which is inconsistent with Flavell et al.'s findings. The difference in our findings could be due to the larger number of trials we collected, which would have introduced more opportunity for fatigue, and this might differentially affect the younger children. One consequence of this finding is that any boost to recall in children younger than 10 that we observe for the point-and-name condition, which always occurred at the end of the session, may be more confidently attributed to the instructed speech intervention rather than to practice effects.

We carried out a Bayesian ANOVA on number of speech behaviors observed during retention with age group as a between-participants factor and condition (delayed recall, point and name) as a within-participants factor. Results tended to confirm the hypothesis that instructions to explicitly name the pictures mediated use of verbalization during retention. The best model (BF = 1.7×10^{43}) included main effects of age group (inclusion favored by a factor of 2.1×10^{13}) and task (inclusion favored by a factor of 8.3×10^{29}). Excluding their interaction was favored by a factor of 99. More verbalizations were observed during point and name ($M = 2.92$, $SD = 2.63$) than during delayed recall ($M = 2.04$, $SD = 2.46$). In both tasks, speech instances increased with age.[4]

We similarly compared average maximum span in delayed recall and point-and-name conditions. Carrying on from the previous analysis of speech behavior, we categorized participants as speech producers or nonproducers; a participant became a speech producer if the participant was observed verbalizing during the delay period of the delayed recall task at least once. By this definition, there were more than 20 speech producers

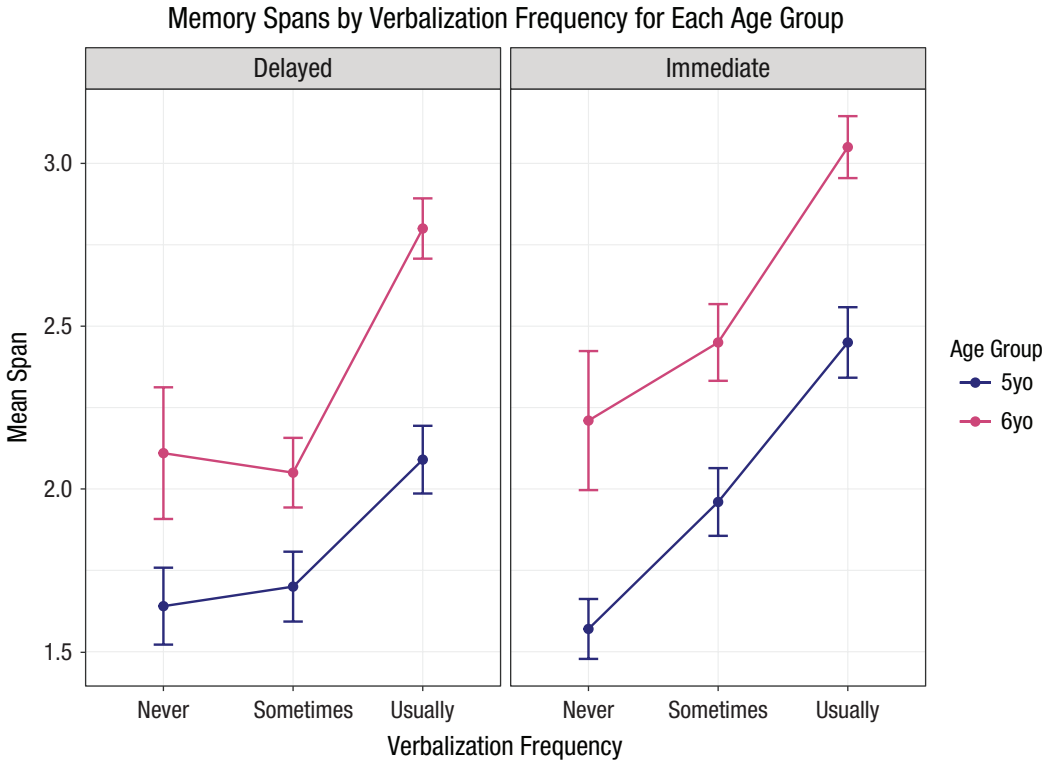

Fig. 5. Memory spans by verbalization frequency and recall task at 5- and 6-years-old. Error bars are within-participants standard errors of the mean (Morey, 2008).

and nonproducers per age group, so we could include all age groups in the analysis. According to a Bayesian ANOVA, the best model included effects of age group, speech-producer classification, task, and all interaction terms, including the three-way interaction,[5] BF = 7.9×10^{142}. Including age group was favored by a factor of 3.4×10^{116}; spans increased as children matured. Including speech-producer classification was favored by a factor of 2.1×10^{14}; participants who were observed verbalizing during the delay period of the delayed recall task remembered more on average in both the delayed recall and point-and-name tasks than participants who were never observed verbalizing during delayed recall. Including the effect of task was favored by a factor of 2.9×10^{7}. The point-and-name procedure did appear to improve span scores. The three-way interaction (inclusion favored only by a factor of 3.87 over the model including all other terms, including each possible two-way interaction) could reflect younger children (particularly 6-year-olds) benefiting from spontaneously verbalizing more than older children and benefiting somewhat more from spontaneous verbalizing than instructed verbalizing (see Fig. 6). Although inclusion of the three-way interaction is only marginally favored over the next most complex model, inclusion of some combination of the two-way interactions is favored by a factor of at least 4.08. This value comes from comparing the model including each possible two-way interaction with the best model including only one two-way interaction (which was the Age Group × Task interaction). The number of speech instances observed correlated with span in both the delayed recall task (r = .38, BF = 6.9×10^{30}) and the point-and-name task (r = .34, BF = 1×10^{25}). Altogether, it seems reasonable to conclude that by age 10, neither the speech-producer status nor instruction to speak during presentation is influencing span very much, but both tendency to speak spontaneously and the point-and-name intervention corresponded to larger spans in younger children.

Discussion

Recent research has cast doubt on the claim that children younger than 7 years old do not spontaneously verbalize when trying to remember picture stimuli. Specifically, the phonological similarity effect has historically been used to argue that only older children recode picture stimuli into phonological labels (Hitch et al., 1988; Hitch, Halliday, et al., 1989; Hitch, Woodin, & Baker, 1989). However, recent work has skeptically considered whether absolute differences between age groups in the size of these effects is strong evidence of

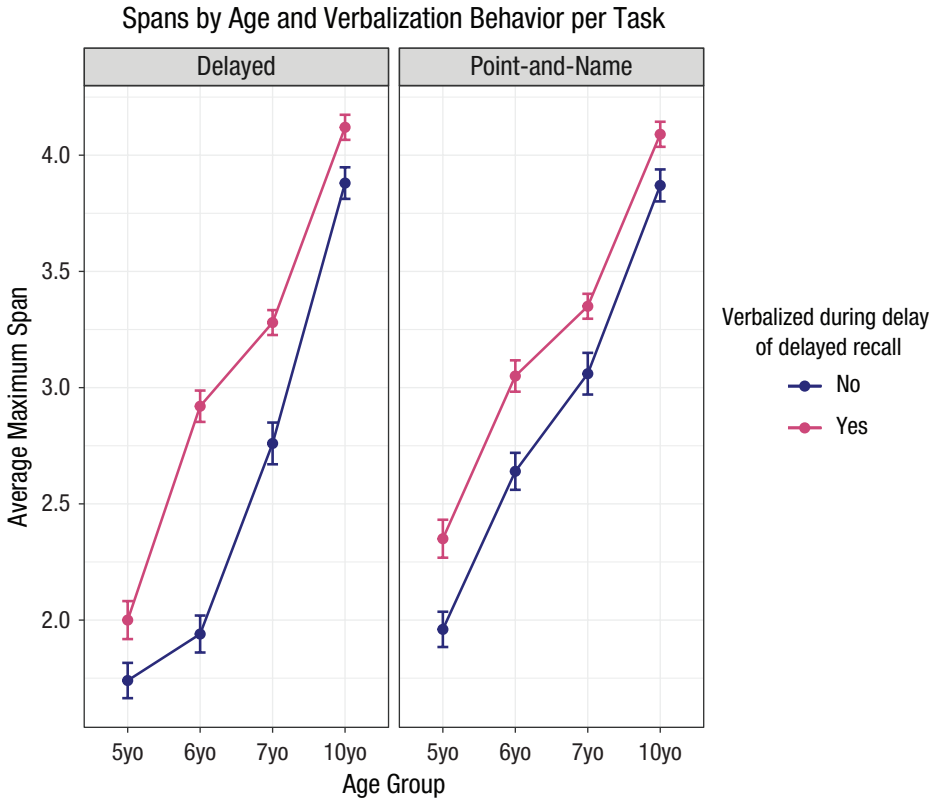

Fig. 6. Mean span in delayed recall and point and name by age and speaker classification. Error bars are within-participants standard errors of the mean (Morey, 2008).

a developmental difference (Jarrold & Citroen, 2013; Jarrold et al., 2015). Note that the phonological similarity effect is an indirect measure of verbalization. Complementary evidence about children's tendencies to verbalize comes from direct observations of verbalization behaviors during serial recall trials, such as those observed in the seminal work of Flavell et al. (1966). Therefore, a replication of Flavell et al.'s original study provides timely evidence needed to assess and update claims about recoding in children: When do they begin verbalizing picture stimuli to be remembered, and when does this verbalization behavior benefit recall accuracy?

Across 17 labs and 977 child participants, this preregistered replication allowed a thorough examination of the development of children's verbalization behaviors and memory performance. The general pattern of developmental changes in verbalization behaviors that Flavell et al. (1966) observed was upheld. However, it was clear from this much larger sample of children that even the youngest children exhibited more speech behavior than would have been expected based on the original work. According to the values in Figure 2, 76% of our 5-year-old group vocalized at least sometimes, compared with 10% of Flavell et al.'s equivalent group. We cannot know for certain the source of this difference. Possibly, our longer sessions with standardized administration of the same span lengths for all participants afforded more opportunities for 5-year-olds to demonstrate verbalization. It could also be the case that cultural expectations around children speaking to adults unbidden differ in our samples or that broader exposure to preschool has increased children's willingness to voluntarily speak up in front of researchers. It is also plausible that the original finding that 5-year-old children rarely spontaneously verbalize would still be obtained with some probability in a random sample but that it is simply a less probable outcome. In any case, observing that spontaneous verbalization is not uncommon in 5-year-olds but nevertheless increases between 5 and 7 years of age provides a critical link between Flavell et al.'s original finding and Jarrold and colleagues' (Jarrold & Citroen, 2013; Jarrold et al., 2015) more contemporary findings that young children display a phonological similarity effect. Therefore, this Registered Replication Report contributes necessary information relevant to the theoretical understanding of the relationship between the developmental changes in memory performance and verbalization behaviors.

Overall, the findings in this Registered Replication Report tend to confirm the notion that children younger than 7 spontaneously apply speech to remembering.

When examining the characterizations of observed verbalizations in Table 2, 167 out of 220 five-year-old children verbalized either sometimes or usually; however, our findings supported the original finding that the youngest children, the 5-year-olds, did verbalize significantly less than the 7- and 10-year-old children. Although the youngest children verbalized less than the older children in the original Flavell et al. (1966) and in the current meta-analysis, we did not confirm that speech continues increasing from 7- to 10-year-olds in lab-level analyses. Speech also did not decrease in 10-year-old children, contrary to expectations based on Vygotsky (1962). Although nearly half of the 7-year-olds Flavell et al. sampled did not overtly verbalize, we observed that most 7-year-olds, and even most 6-year-olds, verbalized at least sometimes. This difference in results would account for our lack of a clear continuous increase in speech after age 7 because by age 7, most of our sample was already overtly verbalizing.

Because of the larger samples in this many-labs project, we were able to perform analyses that Flavell et al. (1966) could not. Our preregistered analysis relating speech behavior to memory span in the 5- and 6-year-old children indicated that qualitative increases in verbalizations resulted in higher memory spans (see Fig. 5). Follow-up exploratory analyses including all ages were also consistent with the prediction that speech improved memory spans (see Fig. 6). Thus, we can complement Flavell et al.'s original article by providing clear evidence linking verbalizations to memory performance and to the success of the point-and-name task as an intervention that served both to increase verbalizations and to improve memory-span performance. Although it has emerged that there is reason for skepticism about the idea of rehearsal benefiting serial recall in adults (Souza & Oberauer, 2018, 2020), our data suggest that increasing verbalizations led to increased memory-span performance in children. This is clearly an area for further research, perhaps including an in-depth look at individual differences in the strategies and behaviors of older children, for whom speech may become less beneficial. We focused very closely on the 5-, 6-, and 7-year-old children to determine the onset of verbalization behaviors and observed considerable variability. Although we observed speech in 5-year-olds, it was less clear that they were aware of using it for remembering than older children were (see self-reported strategies in Table 6), and the benefits of verbalizing were particularly striking for 6-year-old children (see Fig. 6). By age 10, very little benefit with overt verbal labeling was evident, which is consistent with the idea that relationships between overt labeling and memory performance may diminish by adulthood. Moreover, age-related increases in memory were observed in both verbalizers and nonverbalizers (see Fig. 6). The sources of these age-related improvements are unclear, but these improvements are consistent with either an expanding focus of attention (Cowan, 2016) or an increasing refreshing rate (Gaillard et al., 2011), which are both thought to occur independent of rehearsal.

The benefit to memory that accompanies children's rehearsal is surprising only if one assumes that children's rehearsal looks like adult rehearsal from its onset. This apparent qualitative shift from the absence to the presence of rehearsal is consistent with theories that associate rehearsal with the phonological loop; according to these theories, rehearsal emerges around 7 years of age as the phonological loop matures (e.g., Gathercole, 1998). On the surface, the data presented here could support the idea of a spontaneous, all-or-none emergence of rehearsal—provided that the age of emergence be allowed to vary across children. Certainly, the acquisition of language skills, more generally, may also be a prerequisite for rehearsal beyond passive maturation of the phonological loop. In addition to the benefits of robust phonological representations that would be supported by the phonological loop (Baddeley et al., 1998), robust long-term lexical networks contribute to memory-span performance for both children and adults (Edwards et al., 2004; MacDonald & Christiansen, 2002). Moreover, individual differences in phonological robustness, long-term lexical knowledge, and verbalization speed all predict unique variance in adolescents' serial recall of verbalizable pictures similar to those used in the current study (AuBuchon et al., 2019).

Unfortunately, Flavell et al.'s (1966) original design, and thus the design used here, was intended only to describe the ages at which spontaneous verbalization is observed, not to test the mechanisms that contribute to its emergence. However, the efficacy of the point-and-name intervention with children 7 and younger who did not spontaneously speak suggests that the capability to use speech to serve memory is present before children proactively engage it. A qualitative shift might instead reflect this proactive tendency to anticipate how best to fulfill the goals of the task. With development, children may learn to use their newly acquired rehearsal abilities in more effective and complex ways (Guttentag et al., 1987). Cowan et al. (1991) showed that requiring overt naming of list items in a span task increased 4-year-old children's performance in the task only when the list was presented and the naming was to be carried out in a cumulative fashion, a means of simulating cumulative rehearsal in children ordinarily too young to carry it out. Although it remains possible that a specialized rehearsal resource becomes available with developmental maturation, the evidence favoring this is equally compatible with the idea that with maturation, children are increasingly likely to proactively use verbalization to support memory. Future work should

further examine these developmental transitions with the aim of clarifying precisely why verbalization may differentially benefit younger children.

Regional differences appeared minimal even though the ages that children first began formal schooling varied by location. For example, in some countries, the start of formal schooling occurs at a standard age, with little, if any, variability. Although this standard age varied per country, the patterns across the labs were largely similar (see Fig. 1), and any apparent discrepancies were not clearly replicated within a country in which multiple distinct groups from one country participated. It is noteworthy that our samples were primarily from Europe and the United States. Although we do not know whether these findings generalize to children in all cultures, with different languages and different schooling practices, we did observe strong consistency in the samples we obtained despite variability in schooling practices and native language. Evidence suggests that there are changes in children's memory strategies that have a positive impact on performance despite the fact that these strategies are often not taught explicitly by classroom teachers (e.g., Ornstein & Coffman, 2020). There is also evidence from comparisons of children who are chronologically very close in age but differ in their grade in school (Morrison et al., 2019), which further suggests that there are benefits to formal schooling that go above and beyond maturation. One may therefore speculate that the general finding of increased participation of younger children in formal schooling across the regions could have influenced the increased usage of overt verbalization behaviors in younger children.

Although we tried to remain as true as possible to the Flavell et al. (1966) study, we found that the pacing in the point-and-name task (at the rate of 2 s per item) was difficult for many of the youngest children to follow. It is possible that because the original study was run without computer-regulated stimulus presentation, the experimenters slowed their pace when working with younger children. However, the extent of any pacing differences in the original study cannot be known. Therefore, the full effect of the point-and-name task may have been underestimated in the current research because many of the youngest children had difficulty naming stimuli at the appointed pace. Future research could be conducted using a manipulation like the point-and-name procedure with an adaptive pacing that would be individually adjusted for each child. However, despite this limitation, we were able to confirm Flavell et al.'s hunch that verbalization benefited recall, especially for the younger children we sampled.

In summary, the current multilab study replicated the spirit of Flavell et al.'s (1966) original findings: The number of children spontaneously verbalizing increased from our youngest to our older age groups. This finding verifies a premise underlying the past five decades of developmental research that children come to increasingly rely on self-directed speech to regulate behavior and cognition. However, with our larger sample sizes, increased number of trials, and inclusion of 6-year-olds, we were able to make observations unavailable to the original authors. First, very young children were found to verbalize in greater numbers than originally reported, which indicates that the age of emergence of verbalization in service of memory varies individually more than previously concluded. Second, if there is an inflection point in the rate of verbalization development, it appears to occur before 7 years of age. Finally, we were able to verify Flavell et al.'s exploratory analysis and show that verbalizing is related to improvements in memory span.

Transparency

Action Editor: Daniel J. Simons
Editor: Daniel J. Simons
Author Contributions

E. M. Elliott and C. C. Morey jointly generated the idea for the study. E. M. Elliott programmed the study, created the protocol for data collection, and built the OSF page. The different language versions of the program were forward- and back-translated by their respective teams. T. Castelain developed the Spanish version; A. A. Özdoğru developed the Turkish version; E. Vergauwe, B. Valentini, and S. Jeanneret developed the French version; J. R. Lelonkiewicz and D. Crepaldi developed the Italian version; C. K. Tamnes developed the Norwegian version; and S. Poloczek worked with S. Hoehl and J. P. Röer to develop the German version. C. C. Morey generated analysis code to test pilot data and to provide heat maps and other visual representations of the original data for the predata manuscript. A. M. AuBuchon and G. Meissner provided a detailed video for training the researchers to conduct the coding of verbalization behaviors and created the template for recording the verbalization behaviors and strategy coding during the experimental sessions. E. M. Elliott, C. C. Morey, and A. M. AuBuchon cowrote the predata manuscript, with helpful feedback from N. Cowan and C. Jarrold. E. M. Elliott served as the primary corresponding author for all communications with the team of researchers. C. C. Morey created the analysis code for the final analyses, and A. M. AuBuchon reviewed and confirmed the code. E. M. Elliott, C. C. Morey, and A. M. AuBuchon cowrote the final manuscript, with helpful feedback from N. Cowan and C. Jarrold. E. Adams, M. Attwood, B. Bayram, S. Beeler-Duden, T. Y. Blakstvedt, G. Büttner, T. Castelain, S. Cave, D. Crepaldi, E. Fredriksen, B. Glass, A. Graves, D. Guitard, S. Hoehl, A. Hosch, S. Jeanneret, T. N. Joseph, C. Koch, J. R. Lelonkiewicz, G. Lupyan, A. McDonald, G. Meissner, W. Mendenhall, D. Moreau, T. Ostermann, A. A. Özdoğru, F. Padovani, S. Poloczek, J. P. Röer, C. Schonberg, C. K. Tamnes, M. J. Tomasik, B. Valentini, E. Vergauwe, H. Vlach, and M. Voracek contributed to data collection and provided either individual-level

data or aggregated data and input on their individual methods, procedures, and observations. T. Castelain, S. Hoehl, J. R. Lelonkiewicz, D. Moreau, S. Poloczek, C. Schonberg, C. K. Tamnes, B. Valentini, E. Vergauwe, H. Vlach, and M. Voracek reviewed the Stage 2 manuscript. For a complete listing of all authors and their locations, see https://osf.io/ehgav/wiki/home/. All of the lab leaders for each team approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding

This work was supported by the Research Council of Norway (288083 and 223273), UiO:Life Science, the National Institute of General Medical Sciences of the National Institutes of Health under Award P20 GM109023, and the Cardiff University School of Psychology.

Open Practices

Open Data: https://osf.io/vgxkf
Open Materials: https://osf.io/vgxkf
Preregistration: https://osf.io/vgxkf

All data have been made publicly available via OSF and can be accessed at https://osf.io/vgxkf. All materials have been made publicly available via OSF and can be accessed at https://osf.io/vgxkf. The design, analysis plan, and Stage 1 Registered Report manuscript underwent peer review and were preregistered at OSF prior to data collection; these files can be accessed at https://osf.io/vgxkf. All supplemental materials mentioned in the article have been made publicly available via OSF and can be accessed at https://osf.io/pn4rk/. The updates made in between approval of our Stage 1 manuscript and completion of the Stage 2 version can be accessed at https://osf.io/vy39r/. This article has received badges for Open Data, Open Materials, and Preregistration. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

ORCID iD

Emily M. Elliott https://orcid.org/0000-0002-2405-990X

Acknowledgments

We thank the many research assistants who helped with the pilot testing of the protocol and all materials and assisted with data collection. Our OSF page (https://osf.io/pn4rk/) includes additional detail per lab regarding local data collection personnel and procedures. We thank Felix Henninger for consulting with E. M. Elliott in programming the study.

Notes

1. Description of how strategy responses were coded (and calculation of their interrater reliability) was added at Stage 2.
2. Note that these means reflect proportions of trials on which speech was observed, whereas the values in Table 5 reflect the proportions of participants who produced speech in each period.
3. A classical ANOVA with identical structure returned significant effects and absences of significant interactions corresponding to the best fitting model in the Bayesian ANOVA.
4. A classical ANOVA with identical structure also returned significant main effects and a nonsignificant interaction term.
5. A classical ANOVA with identical structure returned a corresponding outcome.

References

Administration for Children and Families. (2019). *History of head start*. Office of Head Start. https://www.acf.hhs.gov/ohs/about/history-of-head-start

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, *141*(5), 931–965. https://doi.org/10.1037/bul0000021

Al-Namlah, A. S., Fernyhough, C., & Meins, E. (2006). Sociocultural influences on the development of verbal mediation: Private speech and phonological recoding in Saudi Arabian and British samples. *Developmental Psychology*, *42*(1), 117–131. https://doi.org/10.1037/0012-1649.42.1.117

AuBuchon, A. M., Pisoni, D. B., & Kronenberger, W. G. (2019). Evaluating pediatric cochlear implant users' encoding, storage, and retrieval strategies in verbal working memory. *Journal of Speech*, *62*(4), 1016–1032. https://doi.org/10.1044/2018_JSLHR-H-18-0201

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review*, *105*(1), 158–173. https://doi.org/10.1037/0033-295X.105.1.158

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2020). *lme4: Linear mixed-effects models using 'Eigen' and S4*. https://CRAN.R-project.org/package=lme4

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLOS ONE*, *5*(5), Article e10773. https://doi.org/10.1371/journal.pone.0010773

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) Phase II: 930 new normative photos. *PLOS ONE*, *9*(9), Article e106953. https://doi.org/10.1371/journal.pone.0106953

Conrad, R. (1971). Chronology of development of covert speech in children. *Developmental Psychology*, *5*(3), 398–405. https://doi.org/10.1037/h0031595

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., Saults, J., Winterowd, C., & Sherk, M. (1991). Enhancement of 4-year-old children's memory span for phonologically similar and dissimilar word lists. *Journal of Experimental Child Psychology*, *51*(1), 30–52. https://doi.org/10.1016/0022-0965(91)90076-5

Department for Education. (2017). *U.K. Government Website Survey of parents*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/669857/SFR73_2017_Text.pdf

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech*, *47*(2), 421–436. https://doi.org/10.1044/1092-4388

Flavell, J. H., Beach, D. R., & Chinsky, J. M. (1966). Spontaneous verbal rehearsal in a memory task as a function of age. *Child Development*, *37*(2), 283–299. https://doi.org/10.2307/1126804

Gaillard, V., Barrouillet, P., Jarrold, C., & Camos, V. (2011). Developmental differences in working memory: Where do they come from? *Journal of Experimental Child Psychology*, *110*(3), 469–479. https://doi.org/10.1016/j.jecp.2011.05.004

Gathercole, S. E. (1998). The development of memory. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *39*(1), 3–27. https://doi.org/10.1017/S0021963097001753

Guttentag, R. E., Ornstein, P. A., & Siemens, L. (1987). Children's spontaneous rehearsal: Transitions in strategy acquisition. *Cognitive Development*, *2*(4), 307–326. https://doi.org/10.1016/S0885-2014(87)80010-2

Henninger, F., Shevchenko, Y., Mertens, U., Kieslich, P. J., & Hilbig, B. E. (2019). *Lab.js: A free, open, online study builder*. PsyArXiv. https://doi.org/10.31234/osf.io/fqr49

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. *Quarterly Journal of Experimental Psychology*, *65*(10), 2033–2053. https://doi.org/10.1080/17470218.2012.672997

Hitch, G., Halliday, M., Dodd, A., & Littler, J. (1989). Development of rehearsal in short-term-memory - Differences between pictorial and spoken stimuli. *British Journal of Developmental Psychology*, *7*, 347–362.

Hitch, G., Halliday, S., Schaafstal, A., & Schraagen, J. (1988). Visual working memory in young children. *Memory & Cognition*, *16*(2), 120–132. https://doi.org/10.3758/BF03213479

Hitch, G., Woodin, M., & Baker, S. (1989). Visual and phonological components of working memory in children. *Memory & Cognition*, *17*(2), 175–185. https://doi.org/10.3758/BF03197067

Hulme, C., & Tordoff, V. (1989). Working memory development: The effects of speech rate, word length, and acoustic similarity on serial recall. *Journal of Experimental Child Psychology*, *47*(1), 72–87. https://doi.org/10.1016/0022-0965(89)90063-5

Jarrold, C., & Citroen, 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

Jarrold, C., Danielsson, H., & Wang, X. (2015). Absolute and proportional measures of potential markers of rehearsal, and their implications for accounts of its development. *Frontiers in Psychology*, *6*, Article 299. https://doi.org/10.3389/fpsyg.2015.00299

Jarrold, C., & Hall, D. (2013). The development of rehearsal in verbal short-term memory. *Child Development Perspectives*, *7*(3), 182–186. https://doi.org/10.1111/cdep.12034

Kail, R., & Park, Y. S. (1994). Processing time, articulation time, and memory span. *Journal of Experimental Child Psychology*, *57*(2), 281–291. https://doi.org/10.1006/jecp.1994.1013

Keeney, T., Cannizzo, S., & Flavell, J. (1967). Spontaneous and induced verbal rehearsal in a recall task. *Child Development*, *38*(4), 953–953. https://doi.org/10.2307/1127095

Kray, J., Eber, J., & Karbach, J. (2008). Verbal self-instructions in task switching: A compensatory tool for action-control deficits in childhood and old age? *Developmental Science*, *11*(2), 223–236. https://doi.org/10.1111/j.1467-7687.2008.00673.x

Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B., & Jensen, S. P. (2020). *lmerTest: Tests in Linear Mixed Effects Models*. https://CRAN.R-project.org/package=lmerTest

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*(4), 1494–1502. https://doi.org/10.3758/s13428-016-0809-y

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review*, *109*(1), 35–54. https://doi.org/10.1037/0033-295X.109.1.35

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorial for Quantitative Methods in Psychology*, *4*, 61–64.

Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). *BayesFactor: Computation of Bayes Factors for common designs*. https://CRAN.R-project.org/package=BayesFactor

Morrison, F. J., Kim, M. H., Connor, C. M., & Grammer, J. K. (2019). The causal impact of schooling on children's development: Lessons for developmental science. *Current Directions in Psychological Science*, *28*(5), 441–449. https://doi.org/10.1177/0963721419855661

Ornstein, P., & Coffman, J. L. (2020). Toward an understanding of the development of skilled remembering: The role of teachers' instructional language. *Current Directions in Psychological Science*, *29*(5), 445–452. https://doi.org/10.1177/0963721420925543

Souza, A. S., & Oberauer, K. (2018). Does articulatory rehearsal help immediate serial recall? *Cognitive Psychology*, *107*, 1–21. https://doi.org/10.1016/j.cogpsych.2018.09.002

Souza, A. S., & Oberauer, K. (2020). No evidence that articulatory rehearsal improves complex span performance. *Journal of Cognition*, *3*(1), Article 11. https://doi.org/10.5334/joc.103

U.S. Department of Education and National Center for Education Statistics. (2016). *Primary early care and education arrangements and achievement at kindergarten entry*. https://nces.ed.gov/pubs2016/2016070.pdf

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(1), 1–48. https://doi.org/10.18637/jss.v036.i03

Vygotsky, L. (1962). *Thought and language* (E. Hanfmann & G. Vakar, Eds.). MIT Press. https://doi.org/10.1037/11193-000