Original Article

QJEP

Quarterly Journal of Experimental Psychology
2021, Vol. 74(12) 2182–2192

DOI: 10.1177/17470218211022635
qjep.sagepub.com

Do production and verification tasks in arithmetic rely on the same cognitive mechanisms? A test using alphabet arithmetic

Jasinta DM Dewi, Jeanne Bagnoud and Catherine Thevenot

Abstract

In this study, 17 adult participants were trained to solve alphabet–arithmetic problems using a production task (e.g., C + 3 = ?). The evolution of their performance across 12 practice sessions was compared with the results obtained in past studies using verification tasks (e.g., is C + 3 = F correct?). We show that, irrespective of the experimental paradigm used, there is no evidence for a shift from counting to retrieval during training. However, and again regardless of the paradigm, problems with the largest addend constitute an exception to the general pattern of results obtained. Contrary to other problems, their answers seem to be deliberately memorised by participants relatively early during training. All in all, we conclude that verification and production tasks lead to similar patterns of results, which can therefore both confidently be used to discuss current theories of learning. Still, deliberate memorization of problems with the largest addend appears earlier and more often in a production than a verification task. This last result is discussed in light of retrieval models.

Keywords

Learning; procedural knowledge; memory; training; verification tasks; arithmetic

Received: 15 March 2021; revised: 9 May 2021; accepted: 13 May 2021

Researchers agree that children start solving additions by counting procedures (e.g., Bagnoud, Dewi, Castel, et al., 2021; Baroody, 1987; Carpenter & Moser, 1984; Groen & Parkman, 1972). However, the way counting strategies evolve with practice is still a matter of debate (see Baroody, 2018; Chen & Campbell, 2018; Thevenot & Barrouillet, 2020, for reviews). Two theoretical views can be contrasted. According to retrieval models, the counting strategies used during childhood are gradually replaced by memory retrieval during the course of development. In adulthood, retrieval is therefore the dominant strategy for all additions involving two single-digit numbers (e.g., Ashcraft, 1982, 1992; Campbell, 1995; Campbell & Oliphant, 1992; Chen & Campbell, 2018; Siegler, 1996). In opposition to this widely accepted traditional view, Baroody (1983, 1984, 1994) argued that simple arithmetic problems could be solved by automated procedures in the form of rules and heuristics. This idea that procedures are still used by experts for very simple addition problems has been taken up recently within the automated counting procedure theory (e.g., Barrouillet & Thevenot, 2013; Fayol & Thevenot, 2012; Mathieu et al., 2016; Uittenhove et al., 2016), according to which the development of strategy in arithmetic could consist in an acceleration of counting procedures until automatization (Thevenot et al., 2016).

Support for the shift from counting to retrieval in the course of learning has been provided by the instance theory of automatization (Logan, 1988). This theory was developed to account for the acquisition of cognitive skills that can be first learnt by algorithm-based procedures. With each instance of learning, a single memory trace associating the stimuli and the response would be created and then stored in long-term memory. Whereas the probability of using algorithm-based procedures is constant in

Institute of Psychology, University of Lausanne, Lausanne, Switzerland

Corresponding author:
Catherine Thevenot, Institute of Psychology, University of Lausanne, Géopolis, Lausanne 1015, Switzerland.
Email: catherine.thevenot@unil.ch

the course of learning, the probability that memory retrieval is used depends on the number of traces in long-term memory. Therefore, with repeated practice, more and more traces will be created and, at some point, the probability of using memory will be higher than of using algorithm. This point corresponds to the shift from procedural-based to memory-based performance, or, in mental addition, from counting to retrieval.

In this framework, the shift from counting to retrieval in mental arithmetic has been studied using the alphabet–arithmetic paradigm, which was conceived to mimic the way children learn additions. In this paradigm, a number addend is added to a letter augend, resulting in a letter answer. For example, A + 5 = F because F is 5 letters away from A. In their seminal work based on a training experiment, Logan and Klapp (1991) asked adults to learn 40 alphabet–arithmetic problems, consisted of 10 letters paired with addends 2, 3, 4, and 5. Half of the participants learnt the first 10 letters of the alphabet (i.e., A to J) and the other half the second 10 letters (i.e., K to T). After the training phase, which lasted 12 days, participants had to work with the other set of letters on the 13th day. Logan and Klapp concluded that the shift of strategy has occurred, because the slope of solution times as a function of addend (hereafter: addend slope) was significant in Session 1, implying the use of counting, but was not significant in Session 12, suggesting the use of memory retrieval (see also, e.g., Chen et al., 2020; Compton & Logan, 1991; Zbrodoff, 1995, 1999). Furthermore, the addend slope during the transfer phase on Day 13 was again significant, implying that there was no transfer and indicating the item specificity of alphabet–arithmetic learning.

However, these classical findings have recently been put into question by Thevenot et al. (2020) who argued that the non-significant addend slope at the end of Logan and Klapp's (1991) training experiment was due uniquely to the decrease in solution times for problems with the largest addend in the study set. Thevenot et al. showed that when these problems were excluded from the analysis, the addend slope was significant until the end of training. This decrease in solution times for problems with the largest addend was observed only in a minority of participants that the authors called the breakers (i.e., 6 out of 19 in Experiment 1 and 7 out of 21 in Experiment 2). For participants who did not show the discontinuity in solution times, the addend slope remained significant for all addends until the end of the training experiment. This constitutes a challenge for the theory of instance automatization (e.g., Logan, 1988) because if the slope is not null at the end of training, the possibility that its reduction and the decrease in solution times during practice are caused by an acceleration of counting procedures cannot be discarded.

Nevertheless, most alphabet–arithmetic studies (e.g., Compton & Logan, 1991; Logan & Klapp, 1991; Thevenot et al., 2020; Zbrodoff, 1995, 1999) were based on a verification task. This can be problematic because counting or memory retrieval could be bypassed in a verification task by the use of plausibility judgements (Reder, 1982). In mental arithmetic, such judgements involve the evaluation of the equation as a whole without exact calculations (e.g., Zbrodoff & Logan, 1990). This includes the situations where the proposed answer deviates largely from the correct answer (e.g., Ashcraft & Battaglia, 1978; de Rammelaere et al., 2001; Zbrodoff & Logan, 1990), when the parity of the proposed answer differs from the parity of the expected result, for example, 4 + 2 = 7, can be easily judged as incorrect because the sum of two even numbers should be an even number (Krueger, 1986; Krueger & Hallford, 1984; Lemaire & Fayol, 1995; Lemaire & Reder, 1999; Masse & Lemaire, 2001), when the proposed answer to a multiplication problem involving 5 does not contain 0 or 5 (Lemaire & Reder, 1999; Masse & Lemaire, 2001), or when the equation is familiar, for example, $3 \times 4 = 12$ can be easily judged as correct because it has been frequently practised (e.g., Lochy et al., 2000).

Another problematic aspect concerning verification tasks is that solution times depend on whether the presented equation is true or false. Indeed, studies on both mental arithmetic (e.g., Ashcraft & Battaglia, 1978; Ashcraft & Fierman, 1982; Ashcraft & Stazyk, 1981; Campbell, 1987; Groen & Parkman, 1972; Parkman & Groen, 1971) and alphabet–arithmetic (e.g., Compton & Logan, 1991; Logan & Klapp, 1991; Thevenot et al., 2020; Zbrodoff, 1999) using a verification task have shown that solution times are faster for true than for false equations. Furthermore, particularly in alphabet arithmetic, solution times in verification tasks depend on whether the proposed answer precedes or succeeds the correct answer (Dewi et al., submitted; Zbrodoff, 1999). For mental arithmetic studies, Ashcraft and Battaglia (1978) explained the difference in solution times between true and false equations by arguing that in a verification task, the evaluation of correctness is executed only after the correct answer has been found. In fact, whereas a production task involves three stages (i.e., encoding of the problem, searching or computing the answer to the problem, and providing the answer), a verification task involves four stages (i.e., the same three stages as in a production task plus the evaluation of the response, wherein the proposed answer in the equation is compared with the correct answer) (Ashcraft, 1982; Ashcraft et al., 1984). In short, verification is production plus comparison. Within a verification task, the evaluation stage depends on the split or distance effect, that is, the rejection times increase with the distance between the correct answer and the proposed answer (Ashcraft & Battaglia, 1978). This is why solution times for true equations are shorter than those for false equations.

Considering that solution times are often regarded as the mirror of the processes implied in problem-solving but that, as already described, solution times in verification and production tasks can differ, Baroody (1984) asserted that solution times in verification tasks are inevitably not

representative of the genuine times it takes to solve a problem in an ecological situation. Furthermore, assuming that memory retrieval is used to solve the problem, Campbell (1987) argued that memory access to the correct answer might be facilitated by the presented answer in the equation. Therefore, according to him, succeeding in a verification task does not necessarily imply that the participant has correctly retrieved the answer. The arguments put forward by Baroody and Campbell make it obvious that verification tasks are generally less ecological than production tasks.

Therefore, the choice between verification and production tasks is crucial when mental arithmetic is investigated, and particularly when the alphabet–arithmetic paradigm is used. Indeed, past conclusions based on the results obtained with this paradigm could be dependent on the overreliance on methodologies based on verification. Moreover, as already stated, this paradigm is supposed to mimic the way children learn additions because both addition and alphabet–arithmetic tasks have to be learnt initially by way of counting and scanning through a familiar sequence. Nevertheless, in real life, children do not learn additions by means of a verification task and, therefore, the results obtained in alphabet–arithmetic verification tasks might not be directly generalizable to addition learning. Thus, by adopting a production task in this article, we aim at verifying that the results from alphabet–arithmetic verification tasks are replicable in a more-ecological production task. Although several studies using production tasks in alphabet arithmetic have already been conducted (Campbell et al., 2016; Chen et al., 2020; Pyke et al., 2019; Pyke & LeFevre, 2011; Rabinowitz & Goldberg, 1995; Rickard, 2004), this article is the only one allowing for a direct comparison between verification and production tasks in alphabet–arithmetic learning. To do so, we designed a training experiment with a production task using exactly the same stimuli as in the verification training reported in Experiment 2 of Thevenot et al. (2020). This material was very similar to the one constructed by Logan and Klapp (1991) except that 8 consecutive letters instead of 10, which were paired with addends from 2 to 6 instead of 2 to 5, were used. Despite these small differences between Logan and Klapp and Thevenot et al.'s material, the central variables, namely, the number of problems to learn and the number of repetitions for each problem, were kept constant across experiments (i.e., 40 problems, and the number of repetitions for each problem, i.e., 12 times per session). Finally, as in Thevenot et al. or Logan and Klapp, the present experiment consisted of 12 learning sessions. This training programme was followed by three transfer sessions for which the results will not be reported in this article.

If the results obtained in an alphabet–arithmetic task using a verification paradigm are replicable in a more ecological production task, we should observe the discontinuity in solution times found in previous alphabet–arithmetic studies (e.g., Compton & Logan, 1991; Logan & Klapp, 1991; Thevenot et al., 2020; Wenger, 1999; Zbrodoff, 1995, 1999) As already explained above, this discontinuity corresponds to a decrease in solution times for problems with the largest addend. Second, when the problems with the largest addend are excluded from the analyses, the residual addend slope should still be significant at the end of training, implying no sign of shift from counting to retrieval.

Method

Participants

Twenty-three students (six females) aged between 18 and 32 years were recruited by means of the student-job websites of the University of Lausanne and the Swiss Federal School of Technology in Lausanne. All participants were native French speakers and they received CHF 190 for their participation.

Written informed consent was obtained for each participant. All procedures performed in this study, involving human participants, have been conducted in compliance with the Swiss Law on Research involving human beings. Because only behavioural data were collected in a non-vulnerable population of adults, the approval of the Canton de Vaud ethic committee was not required. The study was carried out in accordance with the recommendations of the Ethics Committee of the University of Lausanne, following the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Finally, the research protocol that we followed was approved by the Research Committee of the Faculty of Social and Political Sciences of the University of Lausanne.

Material and stimuli

Participants were trained on an alphabet–arithmetic production task (e.g., A + 2 = ?). Half of the participants were assigned to Group 1 and the other half to Group 2. During the learning phase, participants in Group 1 were trained on the first eight letters of the alphabet (i.e., Set 1: letters A to H) and those in Group 2 on the second eight letters (i.e., Set 2: letters I to P). Each letter was paired with addends from two to six, resulting in 40 problems in each set. Each problem was presented four times in a block, and each session comprised three identical blocks of 160 trials. The 160 problems were randomised within each block. Thus, similar to Experiment 1 of Logan and Klapp (1991) and the two experiments of Thevenot et al. (2020), the stimuli contained 40 problems that were presented 12 times in a session.

The experiment was programmed with the DMDX software (Forster & Forster, 2003). Each trial began with a fixation point (*) presented for 500 ms, followed by the problem, which remained on the screen until participants

gave their response orally into the microphone. Then, the problem disappeared from the screen and the screen remained blank for 500 ms until the onset of the next trial.

Participants' responses were recorded in individual .WAV audio files. Solution times, which corresponded to the time elapsed between problem presentation and voice key triggering, were recorded in a separate file. For some trials, the intensity of participants' responses did not reach the threshold at which the voice key could be triggered and the problem remained on the screen until participants repeated their response louder. In such voice-key failure cases, recorded solution times were not correct and they were therefore corrected manually using the CheckVocal software (Protopapas, 2007). The latter software also allows for the verification of the response accuracy.

Procedure

Participants were trained across 12 sessions, corresponding to 12 consecutive working days. They were tested individually in our laboratory, in separate experimental booths. The experimenter was present in the room where the experimental booths are located, but outside the booth. During the weekend, participants were required to do one session of home training consisting of 160 problems presented on paper that they had to solve as quickly and accurately as possible. The 160 problems corresponded to one experimental block.

Results

We excluded the data of four participants either because the accuracy was too low (less than 75% of correct responses for at least two sessions) or because the number of recording errors was too high (i.e., more than 20% for at least three sessions). The data of two other participants were also excluded because they showed non-significant addend slopes in Session 1. This was done because the alphabet–arithmetic task is conceived with the assumption that participants would start the learning process by a counting procedure, which implies significant addend slopes. It is therefore obvious that these two participants had never solved the problem through counting. Thus, the data of 17 participants were included in the analyses, that is, 10 in Group 1 and seven in Group 2.

Accuracy

We first carried out a 12 (Session: 1 to 12) × 5 (Addend: 2 to 6) × 2 (Group: 1 or 2) repeated-measures, mixed-design analysis of variance (ANOVA) on accuracy with Group as a between measure (see Figure 1 for accuracy in Sessions 1, 6, and 12). First of all, an effect of Group was found, $F(1, 15) = 5.71$, $\eta_p^2 = .28$, $p = .03$, with Group 1 participants having lower accuracy (92%) than Group 2 participants (96%). We also found an effect of Addend, $F(4, 60) = 9.55$, $\eta_p^2 = .39$, $p < .001$, with +2 problems being solved with the highest accuracy (95%) and +5 problems with the lowest accuracy (92%). There was also an interaction between Addend and Group, $F(4, 60) = 3.44$, $\eta_p^2 = .19$, $p = .01$. A series of contrasts with Holm correction showed that this interaction was due to +5 and +6 problems being solved with lower accuracy by Group 1 participants (89% and 92% for +5 and +6 problems, respectively) than by Group 2 participants, 95%, $t(15) = 3.12$, $p = .007$ for +5 problems and 96%, $t(15) = 2.67$, $p = .02$, for +6 problems, whereas there was no difference in accuracy between the two groups for problems with addends 2, 3, and 4.

We further found an effect of Session, $F(11, 165) = 5.71$, $\eta_p^2 = .27$, $p < .001$, with accuracy increasing from 84% in Session 1 to 95% in Sessions 6 and 12. This effect did not interact with Group, $F(11, 165) < 1$, but interacted with Addend, $F(44, 660) = 1.91$, $\eta_p^2 = .11$, $p < .001$. A series of contrasts with Holm correction revealed that the interaction was due to the significant linear addend effect in Sessions 1, $t(15) = -3.53$, $p = .01$, 2, $t(15) = -4.12$, $p = .004$, and 3, $t(15) = -2.89$, $p = .04$, with higher accuracy for lower addend. This addend effect disappeared from Session 4 onwards. The three variables did not interact, $F(44, 660) = 1.12$, $p = .28$.

Solution times

To analyse solution times, we removed invalid trials, that is, faulty trials due to technical problems and trials solved incorrectly, which together corresponded to 6.75% of all trials. Furthermore, we removed correct trials with extreme values, which corresponded to 0.33% of the correct trials. The extreme values were defined as trials with solution times shorter than 250 ms as well as trials with solution times larger than the mean for each participant and each session plus 3 times the corresponding standard deviation.

We performed a 12 (Session: 1 to 12) × 5 (Addend: 2 to 6) × 2 (Group: 1 or 2) repeated-measures, mixed-design ANOVA on solution times with Group as the between measure (see Figure 1 for solution times in Session 1, 6, and 12). An effect of Group was found, $F(1, 15) = 6.01$, $\eta_p^2 = .29$, $p = .03$, with Group 1 participants being faster (1,558 ms) than Group 2 participants (2,128 ms). An effect of Addend was also found, $F(4, 60) = 39.76$, $\eta_p^2 = .73$, $p < .001$, with +2 problems being solved the fastest (1,388 ms) and +5 problems the slowest (2,135 ms). Addend and Group did not interact, $F(4, 60) < 1$.

We found an effect of Session, $F(11, 165) = 47.12$, $\eta_p^2 = .76$, $p < .001$, with solution times decreasing from 2,708 ms in Session 1 to 1,749 ms in Session 6 to 1,500 ms in Session 12. This effect interacted with Group, $F(11, 165) = 2.97$, $\eta_p^2 = .17$, $p = .001$, see Figure 1. A series of contrasts with Holm correction revealed that this

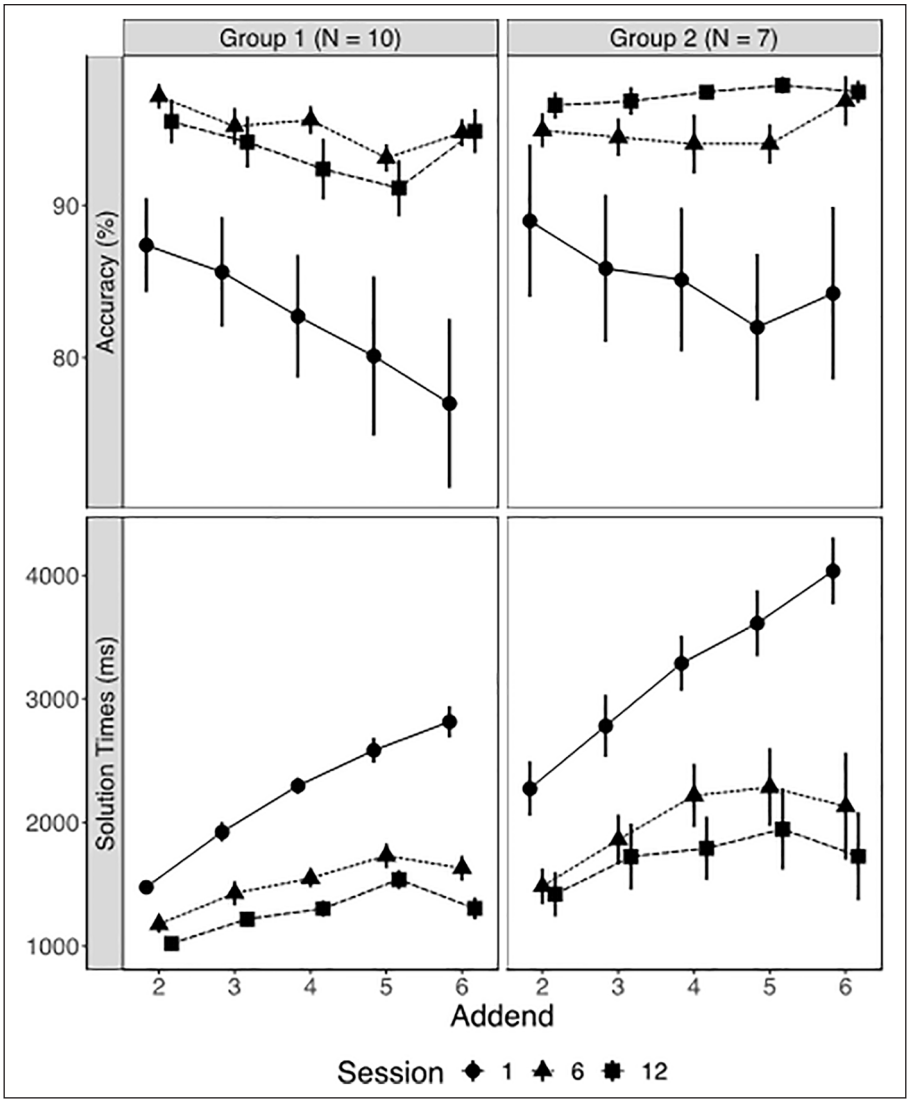

Figure 1. Accuracy and Solution times as a function of addends.
Note. Accuracy (top panels) and solution times (bottom panels) as a function of addends for Sessions 1 (circles, solid line), 6 (triangles, dotted line), and 12 (squares, dashed line) for Group 1 (left) and Group 2 (right) participants. Error bars represent standard errors.

interaction was due to Group 1 participants being faster than Group 2 participants in Session 1, $t(15)=4.47$, $p<.001$; Session 2, $t(15)=2.90$, $p=.01$; and Session 3, $t(15)=2.29$, $p=.04$, but not in other learning sessions.

There was also an interaction between Session and Addend, $F(44, 660)=16.67$, $\eta_p^2=.53$, $p<.001$. The effect of Addend was significant throughout the learning sessions, that is, from $t(15)=14.89$, $p<.001$ in Session 1 to $t(15)=6.14$, $p<.001$, in Session 12. There was no three-way interaction, $F(44, 660)=1.06$, $p=.36$.

As observed when verification tasks are used, Figure 1 shows a discontinuity in solution times in Sessions 6 and 12. In other words, solution times for +6 problems were shorter than for +5 problems. In fact, for both groups of participants, as can be seen in Figure 2, this discontinuity occurred for the first time on average in Session 3. Based on these observations at an individual level, we categorised participants according to whether or not they showed this discontinuity. Two non-breakers did not show a discontinuity in solution times at any point of the experiment. Ten breakers continuously showed a discontinuity starting from one session (i.e., as early as Session 1 and as late as Session 9) until the end of training. Finally, five participants did not show a consistent pattern, that is, they showed a discontinuity in at least one session but the discontinuity disappeared in the following sessions. A χ^2-test of independence revealed that the categorization of participants into breakers, non-breakers, and inconsistent did not depend on the part of the alphabet on which they were trained on, that is, Group 1 or Group 2, $\chi^2(16, N=17)=2.92$, $p=1$.

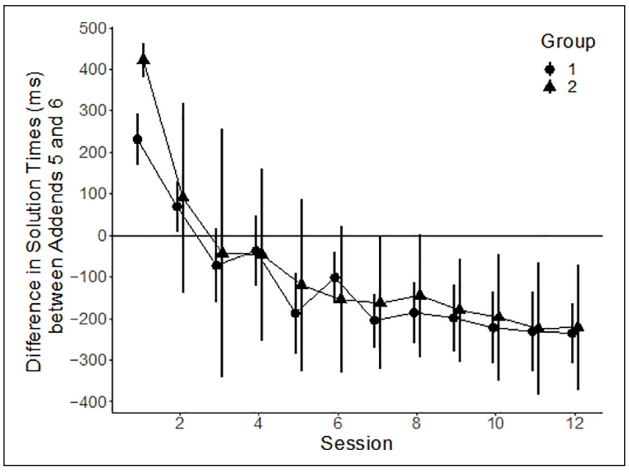

Figure 2. Difference in Solution Times between +5 and +6 problems as a function of sessions.
Note. Difference in solution times between problems with addends 5 and 6 across the 12 sessions. Error bars represent standard errors.

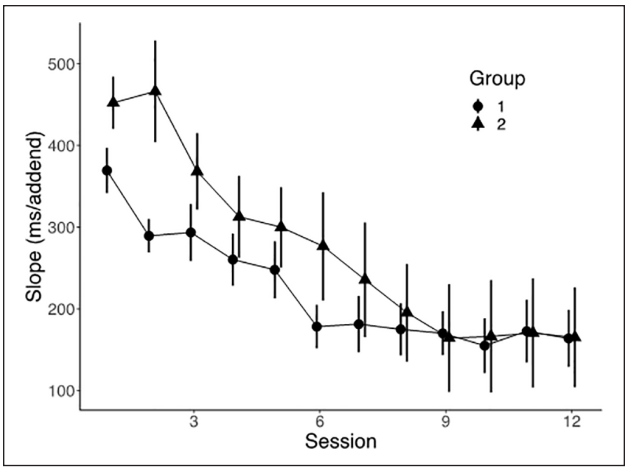

Figure 3. Addend slopes as a function of sessions.
Note. Addend slopes of solution times as a function of sessions for Group 1 (solid circles) and Group 2 (solid triangles), without taking +6 problems into account. Error bars represent standard errors.

Addend slopes

We calculated the addend slopes for each participant and for each session. Because, as just described, the inclusion of problems with the largest addend potentially flattens the addend slopes, we also calculated the addend slopes without +6 problems. Whether +6 problems were included or not, the addend slopes were significantly different from 0 throughout the learning sessions, that is, 385 and 411 ms/addend in Session 1, with and without +6 problems, respectively ($ps < .001$), 146 and 228 ms/addend in Session 6, with and without +6 problems, respectively ($ps < .001$), and 86 ms/addend ($p = .01$) in Session 12 when +6 problems were included and 165 ms/addend when they were not ($p < .001$).

We ran a 12 (Session: 1 to 12) × 2 (data set: with or without +6 problems) × 2 (Group: 1 or 2) repeated-measures, mixed-design ANOVA on addend slope with Group as a between measure. We did not find an effect of Group, $F(1, 15) < 1$, but we found an interaction between Group and Session, $F(11, 165) = 2.12$, $\eta_p^2 = .12$, $p = .02$. A series of contrasts with Holm correction revealed that this interaction was due to a significantly lower addend slope for Group 1 than Group 2, but only in Session 1, $t(15) = 2.30$, $p = .04$; and Session 2, $t(15) = 2.71$, $p = .02$.

More importantly, we found an effect of data set, $F(1, 15) = 35.04$, $\eta_p^2 = .70$, $p < .001$, showing that including +6 problems (172 ms/addend) significantly flattened the addend slope compared with excluding them (247 ms/addend). The effect of data set did not interact with Group, $F(1, 15) < 1$, or Session, $F(11, 165) = 1.25$, $p = .26$. It was significant from Session 1, $t(15) = 2.75$, $p = .01$, to Session 12, $t(15) = 4.29$, $p < .001$. See Figure 3 for a representation of addend slopes without considering +6 problems.

Discussion

In the present research, we investigated whether the results based on an arithmetic verification task are replicated when a more-ecological production task is used. This question is particularly important in the current theoretical context because, as explained in section "Introduction," some assumptions of the instance theory of automatization (Logan, 1988) have been recently called into question using a verification task (e.g., Thevenot et al., 2020). It is therefore central to ensure that previous conclusions of the literature hold in a more-natural paradigm using a production task. To this aim, we replicated Logan and Klapp (1991) and Thevenot et al.'s (2020) experiments using a production paradigm in an alphabet–arithmetic task rather than a verification paradigm, as used in the original experiments.

Exactly as in Thevenot et al. (2020), we found a significant addend slope at the end of the learning phase. Therefore, contrary to Logan and Klapp's (1991) conclusion, the possibility that the decrease in the slopes and the decrease in solution times at the end of an alphabet–arithmetic training is due to an acceleration of procedures rather than a progressive shift from counting to retrieval cannot be discarded. As also observed in previous studies, we found a decrease in solution times for problems involving the largest addend (+6 in the present experiment). Indeed, a discontinuity, or in other words, a drop in solution times was observed for these problems (see Figure 1), which were therefore obviously not processed as the others. The numerous counting steps required to solve problems with the largest addend probably discouraged participants to count. Deliberate memorization of the associations between operands and answers might have therefore been preferred over counting (Logan & Klapp, 1991;

Thevenot et al., 2020). Still, similar to the finding of Thevenot et al., the decrease in solution times, suggesting a deliberate memorization of problems with the largest addend, was not found for all participants.

The observation of a discontinuity in solution times for problems with the largest addend in a production as in a verification task is important because it confirms that end-terms effects in the verification task were not fully responsible to the special processing of these problems. More precisely, in verification task where the false equations are constructed by adding or subtracting a certain quantity to the correct answers (usually 1 or 2 to minimise the split between true and false answers), the answers proposed for the equations involving the largest addend are necessarily underrepresented. For example, in an experiment involving letters ranging from A to H, addends from two to six, and a split of one between the correct and the proposed false answer, the letter O is presented only when H + 6 = O has to be verified. This false equation including the letter O is therefore extremely salient and can be rejected easily. In a production task, such unavoidable statistical irregularities in the material cannot impact the results.

Our result that production and verification tasks lead to the same pattern of results has important implications. It shows that the drop in solution times observed for problems with the largest addend is a generalizable phenomenon. However, in this article using a production task, we found more breakers (i.e., 10 out of 17) than when a verification task was used by Thevenot et al. (i.e., 7 out of 21). Furthermore, the discontinuity observed in this article occurred earlier during practice than in Thevenot et al.'s verification task, that is, Session 3 instead of Session 7. It seems therefore that a production task more strongly elicits deliberate memorization of the associations between the elements of the problems and their answers for problems with the largest addend. An interpretation of this result will be provided later on in section "Discussion."

It is very important to note that the decrease in solution times for problems with the largest addend challenges the instance theory of automatization. As explained by Logan and Klapp (1991), both repeated counting and deliberate memorization should lead to the creation of instances in long-term memory. Indeed, in the framework of the instance theory, what is important for automatization, or in other words for memory retrieval, is the number of traces made and not the way they are created. In this article, as well as in Logan and Klapp (1991) or Thevenot et al. (2020), the number of presentations of problems with the largest addend and the number of presentations of other problems is exactly the same. Therefore, according to the instance theory of automatization, there is no reason for problems with the largest addend to be committed faster to memory than other problems. Thus, the decrease in solution times observed for these problems compared with problems involving an addend immediately inferior to the largest demonstrates that they are not subjected to the principles described in the instance theory of automatization. As a consequence, and at the very least, the slopes calculated in alphabet–arithmetic tasks need to be calculated after the exclusion of problems with the largest addend. As shown in this article using a production task, in Thevenot et al.'s using a verification task, and as estimated from Logan and Klapp (1991) depiction of data, excluding these problems results in a significant addend slope throughout the experiment, from the beginning until the end. As already explained, following Logan and Klapp's rationale that "memory retrieval should produce a slope of zero in the linear function relating reaction time to the magnitude of the digit addend" (p. 180), a decrease in the addend slope across sessions, without its disappearance, is not sufficient to infer a shift from counting to retrieval during the course of training. This invalidates Logan and Klapp's conclusion but we cannot conclude that there was no shift towards retrieval during the experiments. Nevertheless, we can conclude that there is no sign of this shift from the evolution of addend slopes in alphabet–arithmetic tasks.

We have just shown and discussed that the qualitative results we observed concerning our variables of interest are very similar in a production and a verification task. We will now examine whether the results between the two tasks are also quantitatively similar when we consider the other variables that we analysed. The following comparisons are made between the results obtained in the production task reported in this article and the true equations in the verification task reported in Thevenot et al. (2020). As a reminder, the material used in the two experiments is strictly the same. Concerning accuracy, the percentage of errors at the beginning of learning was descriptively higher in the production task (i.e., 8% and 5% in Session 1 for the production and verification tasks, respectively), but this small difference completely disappeared at the end of learning (i.e., 4% in both verification and production tasks). Therefore, the two tasks resulted in very similar error rates. Concerning solution times, it is difficult to make direct comparisons because of the difference in the way solution times are measured in the two tasks, that is, oral response in the production task versus keyboard pressing in the verification task. We can nevertheless compare the decrease in solution times from the first to the last sessions of the learning phase. Again, they were very similar (i.e., a decrease in 49% in the verification task and of 45% in the production task). Finally, concerning the magnitudes of the addend slopes, they were lower in the production than in the verification task. In the first session, the addend slopes were 385 and 411 ms/addend for the production task and 441 and 487 ms/addend for the verification task when +6 problems were included and excluded, respectively. Interestingly, the addend slopes at the end of the verification task were comparable with the addend slopes in the middle of the production task. More precisely, addend

slopes in Session 12 of the verification task were 163 and 236 ms/addend, when +6 problems were included and excluded, respectively, whereas addend slopes in Session 6 of the production task were 146 and 228 ms/addend, when +6 problems were included and excluded, respectively. In Session 12, addend slopes in the production task were much lower than in the verification task, that is, 86 and 165 ms/addend, when +6 problems were included and excluded, respectively.

Concerning the set of problems including problems with the largest addend, an explanation for smaller slopes at the end of training in the production task can be found in light of the results we obtained concerning breakers and non-breakers. As already noted, there were more breakers among the participants in this article using a production task compared with the verification task used by Thevenot et al. (2020), and the breakers in the production task showed the discontinuity in solution times earlier during the learning phase. Furthermore, in the end of the learning session, the difference in solution times between +5 and +6 problems was about 100 ms in the verification task (see Figure 8 of Thevenot et al.) and about 200 ms in the production task (see Figure 2 in this article). All these results show that deliberate memorization of the problems with the largest addend is more prominent in a production than in a verification alphabet–arithmetic task. One possible interpretation is that in a verification task, the false answers that are proposed in half of the trials interfere with the correct answers, hence more difficult associations between the different elements of the problems (e.g., Siegler & Shrager, 1984). Concerning the set of problems without problems with the largest addend, smaller slopes in the production than in the verification task could be due to a higher acceleration of counting procedures in the production task. Alternatively, this difference could be due to more numerous shifts from counting to retrieval in the production than in the verification task. As already mentioned, it is not because such shifts are not evidenced by our results that they never occur. Nevertheless, it is unlikely that further training in a verification task would lead to the same level of performance as in a production task, and a fortiori, would lead to a complete shift from counting to retrieval. Indeed, Thevenot et al. (2020, Experiment 1) ran an alphabet–arithmetic task over 25 instead of 12 sessions and showed that the addend slopes across sessions were always different from 0 (i.e., from Session 1 to Session 25). More crucially for the present point, there was no significant difference in the size of addend slopes between Sessions 12 and 25. The asymptote was therefore reached by Session 12, showing that from this point onwards, there was no further evolution in participants' strategy choices.

To sum up, the overall pattern of results obtained in this study using a production task replicates the results obtained in a verification task. We can therefore conclude that verification and production tasks rely on the same general cognitive mechanisms, at least when the split in the verification task between the correct and the proposed answer is small. It could be interesting to test in future studies whether manipulating the size of the split can affect alphabet–arithmetic tasks (e.g., D + 3 = P). In fact, even if we show here that using a production or a verification task provide similar results, this does not mean that, in the previous literature, all studies using a verification task could have been conducted using a production task and vice versa. Rather, the choice of the task depends on the purpose of the study. If the goal of the researchers is to collect precise and ecological data, then a production task is more appropriate. However, such level of precision is possible only when participants give their response orally and when, to correct for voice-key failures, solution times for each of the oral response are manually adjusted to correspond to the onset of the spectrogram (e.g., Poletti et al., 2021). Despite the precision of such approach, not any questions can be answered directly using a production task. The distance (i.e., split) between the proposed and the correct answers can obviously be manipulated only in a verification task. As already evoked above, this kind of manipulation allowed researchers to discover that when the split is large, individuals do not always engage in a costly solution process leading to the exact answer but can decide that the answer is false on the basis of a plausibility judgement (e.g., Duverne & Lemaire, 2004, 2005; Hinault et al., 2016). The question of whether individuals have interiorised and can use rules such as the parity rule (e.g., the addition of two even numbers cannot result in an odd number) or the multiple-of-five rule (i.e., multiplying a number by 5 necessarily results in an answer ending by a 0 or a 5) can also be easily addressed using verification tasks (e.g., Krueger, 1986; Masse & Lemaire, 2001). Within such design, it is possible to directly observe whether a false equation is rejected quicker when the proposed answer violates the rule than when it does not. It is also possible to infer such rule use in production tasks by comparing solution times on different problems (e.g., involving a 5 or not; Miller et al., 1984), but this approach seems to be more inferential than using a verification task. Finally, verification tasks can sometimes be more appropriate when researchers aim at recording brain activity (e.g., Avancini et al., 2014; Mathieu et al., 2018). Given that arithmetic problems can be mentally represented in a verbal format (Dehaene, 1992), interference between an oral answer and the problem-solving process can be more detrimental to recordings than interference between the solving process and a purer motor task (i.e., pressing a key for decision). Still, to overcome these complications, delayed production or delayed verification tasks can also be used (e.g., Bagnoud, Dewi, & Thevenot, 2021; Didino, 2011).

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was funded by the Chuard-Schmid Foundation of the University of Lausanne.

ORCID iD

Catherine Thevenot https://orcid.org/0000-0002-4997-1882

References

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review*, *2*(3), 213–236. https://doi.org/10.1016/0273-2297(82)90012-0

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition*, *44*(1–2), 75–106. https://doi.org/10.1016/0010-0277(92)90051-I

Ashcraft, M. H., & Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(5), 527–538. https://doi.org/10.1037/0278-7393.4.5.527

Ashcraft, M. H., & Fierman, B. A. (1982). Mental addition in third, fourth, and sixth graders. *Journal of Experimental Child Psychology*, *33*(2), 216–234. https://doi.org/10.1016/0022-0965(82)90017-0

Ashcraft, M. H., Fierman, B. A., & Bartolotta, R. (1984). The production and verification tasks in mental addition: An empirical comparison. *Developmental Review*, *4*(2), 157–170. https://doi.org/10.1016/0273-2297(84)90005-4

Ashcraft, M. H., & Stazyk, E. H. (1981). Mental addition: A test of three verification models. *Memory & Cognition*, *9*(2), 185–196. https://doi.org/10.3758/BF03202334

Avancini, C., Galfano, G., & Szűcs, D. (2014). Dissociation between arithmetic relatedness and distance effects is modulated by task properties: An ERP study comparing explicit vs. implicit arithmetic processing. *Biological Psychology*, *103*, 305–316. https://doi.org/10.1016/j.biopsycho.2014.10.003

Bagnoud, J., Dewi, J., Castel, C., Mathieu, R., & Thevenot, C. (2021). Developmental changes in size effects for simple tie and non-tie addition problems in 6- to 12-year-old children and adults. *Journal of Experimental Child Psychology*, *201*. https://doi.org/10.1016/j.jecp.2020.104987

Bagnoud, J., Dewi, J., & Thevenot, C. (2021). Differences in event-related potential (ERP) responses to small tie, non-tie and 1-problems in addition and multiplication. *Neuropsychologia*, *153*, Article 107771. https://doi.org/10.1016/j.neuropsychologia.2021.107771

Baroody, A. J. (1983). The development of procedural knowledge: An alternative explanation for chronometric trends of mental arithmetic. *Developmental Review*, *3*(2), 225–230. https://doi.org/10.1016/0273-2297(83)90031-X

Baroody, A. J. (1984). A reexamination of mental arithmetic models and data: A reply to Ashcraft. *Developmental Review*, *4*(2), 148–156. https://doi.org/10.1016/0273-2297(84)90004-2

Baroody, A. J. (1987). The development of counting strategies for single-digit addition. *Journal for Research in Mathematics Education*, *18*(2), 141–157. https://doi.org/10.2307/749248

Baroody, A. J. (1994). An evaluation of evidence supporting fact-retrieval models. *Learning and Individual Differences*, *6*(1), 1–36. https://doi.org/10.1016/1041-6080(94)90013-2

Baroody, A. J. (2018). A commentary on Chen and Campbell (2017): Is there a clear case for addition fact recall? *Psychonomic Bulletin & Review*, *25*(6), 2398–2405. https://doi.org/10.3758/s13423-018-1440-y

Barrouillet, P., & Thevenot, C. (2013). On the problem size effect in small addition: Can we really discard any counting-based account? *Cognition*, *128*(1), 35–44. https://doi.org/10.1016/j.cognition.2013.02.018

Campbell, J. I. D. (1987). Production, verification, and priming of multiplication facts. *Memory & Cognition*, *15*(4), 349–364. https://doi.org/10.3758/BF03197037

Campbell, J. I. D. (1995). Mechanisms of single addition and multiplication: A modified network-interference theory and simulation. *Mathematical Cognition*, *1*(2), 121–164.

Campbell, J. I. D., Chen, Y., Allen, K., & Beech, L. (2016). Transfer of training in alphabet arithmetic. *Memory & Cognition*, *44*(8), 1288–1300. https://doi.org/10.3758/s13421-016-0631-x

Campbell, J. I. D., & Oliphant, M. (1992). Representation and retrieval of arithmetic facts: A network-interference model and simulation. In J. I. D. Campbell (Ed.), *The nature and origin of mathematical skills* (pp. 331–364). Elsevier Science.

Carpenter, T., & Moser, J. (1984). The acquisition of addition and subtraction concepts in grades one through three. *Journal for Research in Mathematics Education*, *15*(3), 179–202. https://doi.org/10.2307/748348

Chen, Y., & Campbell, J. I. D. (2018). "Compacted" procedures for adults' simple addition: A review and critique of the evidence. *Psychonomic Bulletin & Review*, *25*(2), 739–753. https://doi.org/10.3758/s13423-017-1328-2

Chen, Y., Orr, A., & Campbell, J. I. D. (2020). What is learned in procedural learning? The case of alphabet arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(6), 1165–1177. https://doi.org/10.1037/xlm0000775

Compton, B. J., & Logan, G. D. (1991). The transition from algorithm to retrieval in memory-based theories of automaticity. *Memory & Cognition*, *19*(2), 151–158. https://doi.org/10.3758/BF03197111

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1–2), 1–42. https://doi.org/10.1016/0010-0277(92)90049-N

de Rammelaere, S., Stuyven, E., & Vandierendonck, A. (2001). Verifying simple arithmetic sums and products: Are the phonological loop and the central executive involved? *Memory & Cognition*, *29*, 267–273. https://doi.org/10.3758/BF03194920

Dewi, J. D. M., Bagnoud, J., & Thevenot, C. (submitted). Automatization through practice: The opportunistic-stopping phenomenon called into question.

Didino, D. (2011). *A study on the representation of the arithmetic facts memory: Cognitively speaking, is the commutativity a property of multiplications and additions?* [Unpublished doctoral dissertation, University of Trento].

Duverne, S., & Lemaire, P. (2004). Age-related differences in arithmetic problem-verification strategies. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences*, *59*(3), P135–P142. https://doi.org/10.1093/geronb/59.3.P135

Duverne, S., & Lemaire, P. (2005). Arithmetic split effects reflect strategy selection: An adult age comparative study in addition comparison and verification tasks. *Canadian Journal of Experimental Psychology / Revue canadienne de psychologie expérimentale*, *59*(4), 262–278. https://doi.org/10.1037/h0087479

Fayol, M., & Thevenot, C. (2012). The use of procedural knowledge in simple addition and subtraction problems. *Cognition*, *123*(3), 392–403. https://doi.org/10.1016/j.cognition.2012.02.008

Forster, K. I., & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, and Computers*, *35*(1), 116–124. https://doi.org/10.3758/BF03195503

Groen, G. J., & Parkman, J. M. (1972). A chronometric analysis of simple addition. *Psychological Review*, *79*(4), 329–343. https://doi.org/10.1037/h0032950

Hinault, T., Tiberghien, K., & Lemaire, P. (2016). Age-related differences in plausibility-checking strategies during arithmetic problem verification tasks. *The Journals of Gerontology: Series B: Psychological Sciences and Social Sciences*, *71*(4), 613–621. https://doi.org/10.1093/geronb/gbu178

Krueger, L. E. (1986). Why $2 \times 2 = 5$ looks so wrong: On the odd-even rule in product verification. *Memory & Cognition*, *14*(2), 141–149. https://doi.org/10.3758/BF03198374

Krueger, L. E., & Hallford, E. W. (1984). Why $2 + 2 = 5$ looks so wrong: On the odd-even rule in sum verification. *Memory & Cognition*, *12*(2), 171–180. https://doi.org/10.3758/BF03198431

Lemaire, P., & Fayol, M. (1995). When plausibility judgments supersede fact retrieval: The example of the odd-even effect on product verification. *Memory & Cognition*, *23*(1), 34–48. https://doi.org/10.3758/BF03210555

Lemaire, P., & Reder, L. (1999). What affects strategy selection in arithmetic? The example of parity and five effects on product verification. *Memory & Cognition*, *27*(2), 365–382. https://doi.org/10.3758/BF03211420

Lochy, A., Seron, X., Delazer, M., & Butterworth, B. (2000). The odd-even effect in multiplication: Parity rule or familiarity with even numbers? *Memory & Cognition*, *28*(3), 358–365. https://doi.org/10.3758/BF03198551

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492–527. https://doi.org/10.1037/0033-295X.95.4.492

Logan, G. D., & Klapp, S. T. (1991). Automatizing alphabet arithmetic: I. Is extended practice necessary to produce automaticity? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(2), 179–195. https://doi.org/10.1037/0278-7393.17.2.179

Masse, C., & Lemaire, P. (2001). Do people combine the parity- and five-rule checking strategies in product verification? *Psychological Research*, *65*, 28–33. https://doi.org/10.1007/s004260000030

Mathieu, R., Epinat-Duclos, J., Sigovan, M., Breton, A., Cheylus, A., Fayol, M., Thevenot, C., & Prado, J. (2018). What's behind a "+" sign? Perceiving an arithmetic operator recruits brain circuits for spatial orienting. *Cerebral Cortex*, *28*(5), 1673–1684. https://doi.org/10.1093/cercor/bhx064

Mathieu, R., Gourjon, A., Couderc, A., Thevenot, C., & Prado, J. (2016). Running the number line: Rapid shifts of attention in single-digit arithmetic. *Cognition*, *146*, 229–239. https://doi.org/10.1016/j.cognition.2015.10.002

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: Comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(1), 46–60. https://doi.org/10.1037/0278-7393.10.1.46

Parkman, J. M., & Groen, G. J. (1971). Temporal aspects of simple addition and comparison. *Journal of Experimental Psychology*, *89*(2), 335–342. https://doi.org/10.1037/h0031198

Poletti, C., Perez, J. F., Houillon, J. C., Prado, J., & Thevenot, C. (2021). Priming effects of arithmetic signs in 10- to 15-year-old children. *The British Journal of Developmental Psychology*. Advance online publication. https://doi.org/10.1111/bjdp.12363

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*(4), 859–862. https://doi.org/10.3758/BF03192979

Pyke, A. A., Bourque, G., & LeFevre, J.-A. (2019). Expediting arithmetic automaticity: Do inefficiency computation methods induce spontaneous testing effects? *Journal of Cognitive Psychology*, *31*(1), 104–115. https://doi.org/10.1080/20445911.2018.1557664

Pyke, A. A., & LeFevre, J.-A. (2011). Calculator use need not undermine direct-access ability: The roles of retrieval, calculation, and calculator use in the acquisition of arithmetic facts. *Journal of Educational Psychology*, *103*(3), 607–616. https://doi.org/10.1037/a0023291

Rabinowitz, M., & Goldberg, N. (1995). Evaluating the structure-process hypothesis. In F. E. Weinert & W. Schneider (Eds.), *Memory performance and competencies. Issues in growth and development* (pp. 225–242). Lawrence Erlbaum.

Reder, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, *89*(3), 250–280. https://psycnet.apa.org/doi/10.1037/0033-295X.89.3.250

Rickard, T. C. (2004). Strategy execution in cognitive skill learning: An item-level test of candidate models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(1), 65–82. https://doi.org/10.1037/0278-7393.30.1.65

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. Oxford University Press.

Siegler, R. S., & Shrager, J. (1984). Strategic choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 229–293). Lawrence Erlbaum.

Thevenot, C., & Barrouillet, P. (2020). Are small additions solved by direct retrieval from memory or automated counting procedures? A rejoinder to Chen and Campbell (2018). *Psychonomic Bulletin & Review*, *27*, 1416–1418. https://doi.org/10.3758/s13423-020-01818-4

Thevenot, C., Barrouillet, P., Castel, C., & Uittenhove, K. (2016). Ten-year-old children strategies in mental addition: A counting model account. *Cognition*, *146*, 48–57. https://doi.org/10.1016/j.cognition.2015.09.003

Thevenot, C., Dewi, J. D. M., Bagnoud, J., Uittenhove, K., & Castel, C. (2020). Scrutinizing patterns of solution times in alphabet-arithmetic tasks favors counting over retrieval models. *Cognition*, *200*, Article 104272. https://doi.org/10.1016/j.cognition.2020.104272

Uittenhove, K., Thevenot, C., & Barrouillet, P. (2016). Fast automated counting procedures in addition problem solving: When are they used and why are they mistaken for retrieval? *Cognition*, *146*, 289–303. https://doi.org/10.1016/j.cognition.2015.10.008

Wenger, M. (1999). On the whats and hows of retrieval in the acquisition of a simple skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(5), 1137–1160. https://doi.org/10.1037/0278-7393.25.5.1137

Zbrodoff, N. J. (1995). Why is 9 + 7 harder than 2 + 3? Strength and interference as explanations of the problem-size effect. *Memory & Cognition*, *23*(6), 689–700. https://doi.org/10.3758/BF03200922

Zbrodoff, N. J. (1999). Effects of counting in alphabet arithmetic: Opportunistic stopping and priming of intermediate steps. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 299–317. https://doi.org/10.1037/0278-7393.25.2.299

Zbrodoff, N. J., & Logan, G. D. (1990). On the relation between production and verification tasks in the psychology of simple arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(1), 83–97. https://doi.org/10.1037/0278-7393.16.1.83