Accepted for publication in *Applied Cognitive Psychology*. This manuscript may differ from the final published version, which will be available from http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1099-0720

Co-thought gestures in children's mental problem solving: Prevalence and effects on subsequent performance

Wim Pouw[1, 2], Tamara van Gog[1,3], Rolf A. Zwaan[1], Shirley Agostinho[2], and Fred Paas[1,2]

[1]Department of Psychology, Education & Child Studies, Erasmus University Rotterdam, Rotterdam, The Netherlands

[2]Early Start Research Institute, University of Wollongong, Wollongong, Australia

[3]Department of Education, Utrecht University, Utrecht, The Netherlands

Author's Information: Correspondence should be addressed to Wim T. J. L. Pouw, Department of Psychology, Education & Child Studies Sciences, Erasmus University Rotterdam, 3000 DR Rotterdam, The Netherlands, phone: +31 10 4089558, email: pouw@fsw.eur.nl, vangog@fsw.eur.nl, zwaan@fsw.eur.nl, shirleya@uow.edu.au, paas@fsw.eur.nl.

Competing Interests: The authors have no conflicting interests to report.

Authors contribution: WP designed and conducted the study. TvG, RZ, SA, FP, have co-designed the study. WP has written the manuscript, with critical revisions provided by TvG, RZ, SA, & FP.

Funding: This research was funded by the Netherlands Organisation for Scientific Research (NWO-PROO, project number: 411-10-908) and supported by Vereniging Trustfonds Erasmus Universiteit Rotterdam (97010.11/14.0798).

Acknowledgment: We would like to thank the Wollongong Science Centre and Planetarium for facilitating this study, Helena Sophia Schmidt for help with scoring data, and Charly Eielts for re-programming the Visual Patterns Test.

Pre-registration and Open Data Statement: The study design, hypotheses, and planned analyses have been pre-registered on the Open Science Framework. Pre-registration form, materials, anonymized data-set, and syntax of statistical analyses can be retrieved from https://osf.io/dreks/.

Abstract

Co-thought gestures are understudied compared to co-speech gestures, yet may provide insight into cognitive functions of gestures that are independent of speech processes. A recent study with adults showed that co-thought gesticulation occurred spontaneously during mental preparation of problem solving. Moreover, co-thought gesturing (either spontaneous or instructed) during mental preparation was effective for subsequent solving of the Tower of Hanoi under conditions of high cognitive load (i.e. when visual working memory capacity was limited and when the task was more difficult). In this pre-registered study we investigated whether co-thought gestures would also spontaneously occur and would aid problem-solving processes in children (N = 74; 8-12 years old) under high load conditions. Although children also spontaneously used co-thought gestures during mental problem solving, this did not aid their subsequent performance when physically solving the problem. If these null-results are on track, co-thought gesture effects may be different in adults and children.

Keywords: co-thought gesture, problem solving, visual working memory capacity, Tower of Hanoi

Introduction

The majority of our hand-gestures emerge in synchrony with speech, usually in service of some communicative goal. Yet, we also gesture when we think in silence, without any intention of communication. These so-called *co-thought gestures,* which may take the form of pointing to locations/objects, or simulating task-relevant actions (e.g., grasping and replacing), are observed in a variety of non-communicative tasks, such as mental rotation, or remembering a route (e.g., Chu & Kita, 2008; Logan, Lowrie, & Diezmann, 2014). Evidence suggests that such co-thought gestures are not merely epiphenomenal to thinking, because problem solvers' performance has been shown to improve from gesturing (as opposed to not gesturing or being prohibited from gesturing; e.g., Chu & Kita, 2011; Pouw, Eielts, Van Gog, Zwaan, & Paas, under review; So, Shum, & Wong, 2015). Yet, the cognitive function of co-thought gestures is understudied relative to co-speech gestures. As a consequence, it is still unclear when and why co-thought gestures are produced, and whether and how they support cognitive processes.

According to recent evidence, co-thought gestures and co-speech gestures may have a common cognitive origin (Chu & Kita, 2015). That is, the rate with which co-thought gestures are spontaneously produced in a silent mental rotation task corresponds (within participants) with the rate with which co-speech gestures are elicited in the same task. Moreover, when objects are seen as more difficult to manually manipulate, both co-speech and co-thought gestures are less likely to spontaneously emerge (as opposed to objects with a more manipulable surface). Such findings hint at a possible common cognitive origin of co-thought and co-speech gestures that are not directly tied to speech processes or communicative intent (cf. McNeill, 2008). Rather, an action-generation system may underlie such gestures observed in mental rotation tasks, which is

sensitive to affordances solicited by the objects that is thought or spoken about (Hostetter & Alibali, 2008).

In addition to the issue of which cognitive processes (e.g., action readiness) play a causal role in co-thought gesture production, there is the issue of whether, and if so how, co-thought gestures play a causal role in cognitive processes (Chu & Kita, 2011; Kita, Alibali, & Chu, in press; Pouw, de Nooijer, Van Gog, Zwaan, & Paas, 2014). That co-thought gestures affect cognitive processes is supported by a handful of studies (e.g., Chu & Kita, 2011; Hegarty, Mayer, Kriz, & Keehner, 2005; Logan et al., 2014; Pouw, Eielts, Van Gog, Zwaan, & Paas, under review; Schwartz & Black, 1999; So, Shum, & Wong, 2015). For example, mental rotation co-thought gestures are found to improve mental rotation performance (Chu & Kita, 2011). For example, judging whether a cup of a particular size will spill water can be improved when physically enacting a pouring movement through silent gesture (as opposed to mental inference alone; Schwartz & Black, 1999). Arguably, enacting a grasping movement allows bringing forth procedural knowledge from previous experience, which improved performance in this task. Route learning can be improved when participants rehearse the route on a road map with silent tracing gestures (as opposed to rehearsing it verbally; So et al., 2015). In sum, co-thought gestures are not merely epiphenomenal to cognitive processing; they seem to directly support those very processes.

With regard to when and why co-thought gestures are produced, evidence seems to suggest that - like co-speech gestures - they are used to support problem solving especially when internal cognitive resources are taxed (Pouw et al. de Nooijer, van Gog, Zwaan, & Paas, 2014). That is, compared to not gesturing, co-speech and co-thought gestures are more likely to arise when the problem at hand is more difficult (e.g., Chu & Kita, 2008; Hostetter, Alibali, & Kita,

2007). Such gestures seem to provide additional cognitive resources when engaging in cognitively demanding dual-tasks or when internal cognitive resources such as working memory capacity is low (e.g., Goldin-Meadow, Nusbaum, Kelly, & Wagner, 2001; Marstaller & Burianová, 2013).

In a recent study with adults, the idea was tested that co-thought gestures occur when internal cognitive resources are taxed, either induced by the difficulty of the task, or by limited cognitive capacities (Pouw et al., under review). Participants had to mentally prepare for physically solving the Tower of Hanoi at different levels of complexity (3-disc and 4-disc), and were allowed to gesture (i.e. they were not instructed to, but also not prohibited from gesturing) or instructed to gesture. Participants' subsequent problem solving performance (i.e., solving speed) improved when they gestured during mental preparation (either spontaneously or instructed), but only when the TOH task was more difficult, and when they had lower visual working memory capacity (determined via a Visual Patterns Test, explained below). In a subsequent study (Pouw, Mavilidi, van Gog, & Paas, 2016), it was found that adult participants (age 24-50 yrs) who were instructed to gesture during mental problem solving of the TOH, and especially those with a lower visual working memory capacity, were likely to reduce their eye movements during mental problem solving (as indicated by a drop in fixations directed at the Tower of Hanoi display when gesturing vs. not gesturing). This suggests that gesturing changes problem solving processes (as indicated by changes in eye-gaze behavior) as a function of internal cognitive resources available to solve the task (visual working memory capacity) (see Pouw, Mavilidi, van Gog, & Paas, 2016 for further discussion). Internal cognitive resources are thus to be taken into account when assessing effects of gesture on problem solving.

The aim of the current study is to investigate whether our findings with adults, showing that co-thought gestures are effective for fostering subsequent problem solving when cognitive load is high, would be replicated in children (we pre-registered this study). It is important to attempt to replicate these findings in a younger age sample (8-12 years), because children's visual working memory capacity is likely still in development (Gathercole, Pickering, Ambridge, & Wearing, 2004), and as such co-thought gestures could prove to be especially effective for this age sample (see also Paas & Sweller, 2012). Moreover, research on children's spontaneous adoption of co-thought gestures, and the effects on problem solving performance, is particularly scarce (for an exception, see Delgado, Gómez, & Sarriá, 2011).

In the current study, we assessed children's visual working memory capacity and then had them solve two consecutive trials of the Tower of Hanoi problem that differed in complexity, and hence, working memory load (3-disc and 4-disc). Before each problem-solving phase children were required to mentally prepare the task, by thinking through the solution in silence - just prior to physically solving the particular TOH trial. One third of the children were instructed to gesture during mental problem solving, whilst the rest were allowed (but not instructed) to gesture, and we assessed whether they spontaneously gestured (spontaneous gesture group), or not (no gesture group). We hypothesized that gesture prevalence during mental problem solving in children (either when children were instructed to gesture, or when spontaneously gesturing) would positively affect actual problem-solving performance (solving speed and solving steps) in all trials (TOH 3-disc and TOH 4-disc) as compared to natural non-gesturing. Furthermore, this effect would be more pronounced on the more complex task (TOH4) and for children with lower visual working memory capacity. We will further refer to this hypothesis as the gesture effect hypothesis.

Method

Participants & Design

This study was approved by the Human Research Committee of the University of Wollongong. Children between 8 and 12 years old were recruited for participation during their visit to a local Science Centre & Planetarium. This is a facility where children can learn about science, engineering and technology through interactive exhibits. The Tower of Hanoi is one of these interactive exhibits and thus was considered an ideal context for this study. The initial plan was to recruit 100 participants, but within the available time at the science centre we recruited 74 participants. We had to exclude 3 participants due to video camera failure, and an additional 3 participants, because they did not fit the age-requirement (younger than 8 or older than 12) years[1]. Additionally, 2 participants were excluded because they failed to comply with the task procedures (e.g., did not engage in the mental solving task), as observed by both the first and second coder of the video data (e.g., being continuously distracted, talking continuously during the mental solving phase). The final sample consisted of 66 participants (38 boys (57.6%), 28 girls (42.2%), M_{age} =9.83, SD = 1.20).

As stated in the pre-registration report, we assigned one third of participants to the gesture instruction condition (after exclusion: N = 23, 13 boys, 10 girls, M_{age} = 9.96, SD = 1.21), and two thirds to the no gesture instruction condition (after exclusion: N = 43, 15 boys, 28 girls, M_{age} = 9.75, SD = 1.21), This was done because participants in the no gesture instruction condition would later be subdivided in the analyses, depending on whether they gestured spontaneously during the mental problem-solving phase (spontaneous gesture group) or not (no gesture group). For the 3-disc Tower of Hanoi (TOH 3) the spontaneous gesture group included

[1] Although parents were briefed about the age-requirement, due to some miscommunication, some parents enrolled younger/older children for the study.

13 participants, vs. 30 in the no gesture group and 23 in the instructed gesture group; in the 4-disc Tower of Hanoi (TOH 4) the spontaneous gesture group included 19 participants, vs. 24 in the no gesture group and 23 in the instructed gesture group.

Materials & Measures

Visual working memory capacity: Visual Patterns Test. The Visual Patterns Test (VPT; Della Sala, Gray, Baddeley, & Wilson, 1997) was used as a measure of visual working memory capacity. We used an adaptation of the VPT developed by Chu and colleagues (2013). Participants were shown a matrix, in various patterns, in which half of the squares were coloured black. Each pattern was displayed for 3 seconds, after which all the squares turned white. Subsequently, participants indicated by mouse-clicking on an empty grid the previous pattern of black squares. This was the adaptation made to simplify the task for children. The previous version of the task (administered to adults) required participants to verbally recall the pattern by naming letters that are assigned to the squares. Children could select and deselect squares and continue to the next trial when they clicked on 'next' button. The VPT consists of 25 trials, with blocks of five trials per difficulty level (which increased from seven to 11 black squares). Difficulty progressed with increasing trial number. Before the start of the task participants were provided with two practice trials (of three and four black squares, respectively). If participants failed to recall one or more of the black squares, the trial was automatically scored as an incorrect response. After five consecutive incorrect responses within one block of trials the task automatically stopped. Performance scores were calculated as the proportion of correct responses out of all trials (i.e., number of correct trials/total number of trials [i.e., 25]).

Tower of Hanoi. The TOH consisted of a structure with a rectangular base with three evenly spaced pegs mounted on top. The task unfolded with a number of differently sized discs (practice trial: two discs, first trial: 3-discs, second trial: 4-discs) placed on the left-most peg; discs were placed decreasing in size from bottom to top. The discs were to be replaced during the problem-solving process onto the right-most peg in the same order (decreasing in size from bottom to top) while following the rules (see procedure). Identical to the procedure used in Pouw et al. (under review), Participants engaged in mental preparation for 2.5 minutes and then physically solved the problem when they took longer than five minutes for physical solving, the trial was aborted.

Gesture. Each participant's mental preparation phases for the 3-disc and 4-disc trials were coded for prevalence (no gesture vs. gesture) and type (pointing vs. iconic) of gesture. However, since there were virtually no iconic gestures observed (there were two exceptions where children momentarily gestured as-if grasping the discs next to performing pointing gestures) we only examined pointing gestures (see Figure 1 for examples of observed gestures). We counted the number of pointing movements per participant as a measure of gesture frequency in such a way that each rest point after a pointing gesture (either whole hand –or finger-pointing) is considered as one instance and as a trial wherein gesture is prevalent (gesture prevalence). It should be noted that participants were able to, and sometimes did touch the rectangular base of the TOH (or just a place on the table) thereby "marking" a place through pointing gesture instead of pointing only in the air (these pointing-touch gestures were counted as pointing gestures). Additionally, there were four participants who did not point during mental preparation but when they asked a question to the experimenter during this session they did use pointing gestures. These gestures were not considered as gestures during mental preparation, as these concerned co-

speech communicative gestures[2]. An independent coder counted all the gesture instances in the sample and the first author recounted gestures of 15.15% of the participants who gestured to check for interrater reliability (interrater reliability was high, $r = .992$, $p < .001$).

Self-report Measures: Mental effort, difficulty, interest. We obtained an indication of experienced mental effort, perceived difficulty, and experienced interest after the 3-disc as well as the 4-disc TOH problem-solving trial. These self-reports (for a discussion on self-report and cognitive load see Paas, Tuovinen, Tabbers, & Gerven, 2003) were verbally reported and administered by the experimenter on a 5-point rating scale: "How hard did you need to think to solve the task? (mental effort; 1 = 'not hard', to 5 = 'very hard'), "How difficult did you find this task" (difficulty; 1 = 'not difficult', to 5 = 'very difficult'), and "How interesting did you find this task" (interest; 1 = 'not interesting', to 5 = 'very interesting').

Prior experience. Before the start of the TOH task participants were asked whether they had played the Tower of Hanoi before ('no' responses: 83.8 %, 'yes' responses: 16.2%; yes responses were equally distributed among the no gesture instruction condition 16.6 % 'yes' and the gesture instruction condition 15.9% 'yes')[3]. If participants reported yes, we informally assessed whether they were skilled players or merely remembered having played the game. None of the participants reported extensive experience with the task.

Procedure

The current study was conducted at the local science centre. Visiting parents/caregivers (from hereon: caregivers) and their children were invited to participate in a 30 minute study

[2] Note that exploratory analyses revealed that this decision does not alter our results in any significant way.
[3] Note that exploratory analyses revealed that when excluding participants who had played the game in the past the results were not altered in any significant way.

about problem solving. Caregivers were asked whether (one of) their children was between 8-12 years old, and were further informed about the nature of the study and given an information sheet and consent form. Additionally, caregivers were recruited and informed about the study via email through the member's list of the centre, upon which an appointment could be made to conduct the study at the science centre. As a reward for their participation, children were given a discount at the science centre's shop of 6 Australian Dollars.

When caregivers and children agreed on participation they were directed to a more quiet space with minimal background noise, and no other visitors present. However, some background noise of other visitors and exhibitions was unavoidable. The caregivers were allowed to be (and often were) present during the experiment, and were seated behind the child and were asked to not help the child or otherwise intervene during the study in any way. Children were seated at a desk, with a computer, a video camera and the TOH present and were informed about the tasks. Additionally, children were told that there were no right or wrong answers, that results would not be shared with their caregivers, and that they were able at all times to abort the study.

In the first phase of the study children performed the VPT. Children would first read the instruction, upon which the experimenter verbally repeated it, and they proceeded with the VPT-practice trials. If children made a mistake during the practice trials, the experimenter would explain what went wrong and in such cases children re-did the practice trials until successfully completed. Children were then told that a longer series of trials would commence and that they would do that on their own for about 5 minutes, after which they would proceed to the next task.

Subsequently, children engaged in the TOH problem-solving trials. The experimenter explained the nature of the task with a practice example of 2-discs. The children were told that each trial would involve the experimenter putting a tower of discs on the left-most peg from

large to small (i.e., smaller discs on top). They would need to solve the task by replacing the discs to the outer most right peg while taking two rules into consideration. Firstly, only one disc can be moved at a time. Secondly, only smaller discs can be put upon larger discs, not the other way around, and discs of any size could always be put on empty pegs. Additionally, the experimenter demonstrated that you could always move discs back to the original peg if needed, as long as the rules were not violated. Children then performed the practice TOH 2-disc trial. When children were unable to solve the task (which did not happen often), the experimenter repeated the instruction and children redid the practice trial until successfully completed. Subsequently, children were presented with a TOH 3-disc problem, and the rules of the task were repeated. They were told they would have to solve the puzzle as fast and accurately as possible, but before doing so they could prepare their solution for 2.5 minutes (150 seconds) without physically interacting with the TOH. This was called the mental preparation phase, and children were informed that thinking out the moves before the actual task could help them understand the problem. Participants in the gesture instructed condition were explicitly instructed to think with their hands using pointing gestures (as demonstrated by the experimenter who performed several pointing movements in the air directed at the TOH apparatus) during this mental preparation phase. Participants in the no gesture instruction condition were not given this additional instruction. During the mental preparation phase(s), the participant's hand gestures that (spontaneously) emerged during thinking out the solution for 2.5 minutes were video-recorded. Note that sometimes children asked a question during the mental preparation phase. If the question concerned the task rules, the experimenter would give the answer, and instruct the child to (mentally) work again on the problem. Directly after each preparation phase participants solved the task by physically manipulating the TOH (which was also video-recorded). This

procedure, that is, a 2.5 minute mental preparation phase followed by actual problem solving, was performed first with the TOH 3-disc, and then with the more difficult TOH 4-disc.

During the actual solving of the TOH tasks, participants had to solve the task within 5 minutes (they were not informed about this time constraint to avoid them experiencing pressure). If they were not able to solve the task in time (which was sometimes the case on the TOH 4-disc), the trial was aborted, but the experimenter would give pointers to the child on how to finish the task. Although the children mastered the rules of the TOH quite easily, they did sometimes make a mistake with regard to one of the rules, and the experimenter would instruct the children to look again to make sure that children recognized and corrected the violated rule. This always led to self-correction upon which the child could proceed further.

After the TOH 3-disc and 4-disc procedure children were given a final task for exploratory purposes; it involved another 4-disc task with the rules inversed preceded by a self-explanation phase. Data on this task will not be reported here (as registered in the pre-registration report). Finally, children were informed about the nature of the study, thanked and awarded with a voucher for their participation, and contact details of caregivers were gathered for future communication of the results of the study (reported at group level, never on individual children).

Outliers

As mentioned in the pre-registered report, it is likely that the current sample will have considerable variability in visual working memory capacity and problem-solving competence. We used a similar procedure to control for extreme outliers as used by Chu, Meyer, Foulkes, and Kita (2014 see also Miyake et al., 2000). For each independent and dependent variable included in the regression analysis (VPT, TOH3 & TOH4 solving speed, TOH3 & TOH4 number of solving steps) any value laying 3 standard deviations under (or above) the mean will be set to

exactly 3SD under (or above) the mean. This trimming procedure allows us to prevent loss of data with extreme values, without dramatically biasing the results.

Results

The results are divided into three parts. First, relevant descriptive statistics and correlations are reported. Subsequently we report the confirmatory analysis according to plan (see pre-registration report).

Descriptive Statistics and Correlations

TOH Problem-Solving Performance. All participants were able to solve the TOH 3-disc within 5 minutes. Mean solving time was 46.33 seconds (s) (*SD* = 0.72 s, observed range 9-272 s; one outlier > 3SD above the mean was observed and replaced as reported above; new range: 9-173). The mean number of solving steps for TOH 3-disc was 10.02 (*SD* = 5.16, observed range: 7-26; one outlier > 3SD above the mean was observed and replaced; new range: 7-25). Furthermore, 62.1% (41/66) of the children solved the TOH 3-disc in the fastest way possible (i.e., 7 steps). [4]

Seven participants (10.6%) were not able to solve the TOH 4-disc within 5 minutes, so for these participants no score was obtained. Mean solving time was 111.02 s (*SD* = 65.84, observed range: 26-278 s; no outliers). The mean number of solving steps for the TOH 4-disc was 31.83 (*SD* = 13.22, observed range: 15-74; one outlier > 3SD above the mean was observed

[4] It is valuable to compare the current results with the adult sample (Pouw et al., under review). The analyses with adults consisted of 73 Dutch University students that were tested in the lab (no instruction condition = 38, instruction condition = 35). In adults the mean solving time for the TOH3 disc was 20.52 seconds (*SD* = 16.00), with 82.2% of the adults solving the problem in the minimum number of steps.

and replaced; new range: 15-63). Only 4.5% (3/58) of the participants were able to solve the task in the minimal number of moves (i.e., 15 steps).[5]

VPT. The mean score on the VPT was .42 (*SD* = .20, observed range: .04-.92; no outliers were observed).

Gesture Production. In the no gesture instruction condition during the TOH 3-disc, 30.95% (13/42)[6] of the participants spontaneously gestured during the mental preparation phase with a mean gesture frequency of 48.85 (*SD* = 13.57; observed range: 2-166). In the gesture instruction condition, every participant gestured, and the mean gesture frequency was 77.13 (*SD* = 29.37, observed range: 27-134). The difference between spontaneous gesturers and instructed gesturers in mean gesture frequency in the TOH 3-dics was statistically significant, t (34) = -2.17, p = .037.

On the TOH 4-disc, 44.19% (19/43) of the participants spontaneously gestured during the mental preparation phase in the no gesture instruction condition, with a mean gesture frequency of 43.211(*SD* = 34.82, observed range 7-123). In the instructed gesture condition, all participants gestured, and the mean gesture frequency was 72.09 (*SD* = 34.79, observed range 35-158). The difference between spontaneous gesturers and instructed gesturers in mean gesture frequency in the TOH 4-dics was statistically significant, t (40) = -2.851, p = .007.

[5] In adults (Pouw et al., under review) the mean solving time for the TOH 4-disc was 84.61 seconds (*SD* = 57.83), with 37.7% of the adults solving the problem in the minimum number of steps.

[6] Note that analysis including gesture and performance on the TOH3, one additional participant was excluded because of a camera failure during this trial (but not the subsequent TOH4 trial).

Note, that similar to the adult sample (Pouw et al., under review) in the current study only pointing gestures were produced, as opposed to iconic gestures that mimic grasping actions on the Tower of Hanoi (e.g., Beilock & Goldin-Meadow, 2010).

Gesture Production Relative to Adults. Figure 2 provides the graphical presentation of the gesture production data presented above, with an additional comparison of the gesture production rate (and percentage of spontaneous gesturers) of our previous study with the adult sample (Pouw et al., under review) . Note that exactly the same procedure was used, thus a direct comparison is informative. Informal inspection of Figure 2 shows that the likelihood that spontaneous gestures are adopted in the no instruction group seems to be comparable across age sample (i.e. there is only a 1.64% [TOH3] and 2.09% [TOH4] difference in the likelihood that participants spontaneously gesture). Additionally, it seems that children produced gestures with a frequency equivalent to adults on the 4-disc Tower of Hanoi, but a higher gesture rate on the 3-disc Tower of Hanoi. We will return to the finding that gesture rates are equivalent for children and adults across the board in the discussion.

- Insert Figure 2 here -

Correlations between Problem-Solving Performance, VPT, and Gesture. Table 1 shows the overall correlations and correlations per group (no gesture-, spontaneous gesture-, and instructed gesture group) between problem-solving performance on the TOH 3-disc and 4-disc, VPT, and gesture frequency. As can be seen, overall, VPT is not a strong predictor for problem solving performance, showing only a significant correlation with solving steps on the TOH 4-disc (p = .033), such that participants with higher VPT scores solved the problem in a lower number of steps. Not surprisingly, solving speed and number of solving steps needed to solve the TOH tasks were highly correlated. A more interesting finding of the correlation analyses is that

for participants in the instructed gesture group, the number of gestures performed (i.e., gesture frequency) was highly correlated with problem solving performance on TOH 3-disc, such that higher gesture frequencies resulted in faster solving times (p = .001) and a lower number of steps needed to solve the task (p = .019). However, surprisingly, this was not the case for those participants who spontaneously gestured. Furthermore, this effect of instructed gesture frequency on problem-solving performance was not replicated for the more difficult TOH 4-disc task.

- Insert Table 1 about here -

Self-Report Data. Overall means for the TOH 3-disc and 4-disc are reported in Table 2. As can be expected the TOH 4-disc was generally reported to be more difficult, t (65) = -9.236, p < .001 (paired t-test), and to require more mental effort, t (65) = -8.443, p < .001 than the TOH 3-disc. The children found the TOH 4-disc more interesting than the TOH-3 disc, t (65) = -2.345, p = .022.

-Insert Table 2 about here -

Confirmatory Hypothesis Testing

The following analyses were registered in the pre-registration report.

Hypothesis 1. Our first hypothesis was that after controlling for age, visual working memory (VWM) capacity would be positively related to problem-solving performance (i.e., that higher VWM would result in faster TOH solving speed and fewer steps needed to solve the task). We assessed whether this was the case with stepwise multiple regressions analyses, entering age and VWM as predictors for performance[7]. As can be seen in Table 3, no significant effects of

[7] Note, that Age and VPT were correlated (r = .56, p = .000), and issues of multicollinearity could arise. However, since there were no strong relations with the dependent variable (performance) multicollinearity statistics did not

age or VPT score on performance (i.e., solving speed and solving steps) were obtained on either the TOH 3-disc or TOH 4-disc. This was unexpected and suggests that visual working memory capacity and age were not strong predictors for performance on the problem-solving task.

Hypothesis 2. Our second hypothesis was that after controlling for age, gesture prevalence (instructed and spontaneous) during mental problem solving would positively affect actual problem-solving performance (i.e., faster solving speed and less solving steps) in both trials (TOH 3-disc and TOH 4-disc) as compared to natural non-gesturing, but that this effect would be more pronounced on the most complex task (TOH4) and for participants with a relatively lower visual working memory capacity. We analysed whether this was the case with two multiple stepwise regression analyses per DV. For each DV (3-disc and 4-disc solving speed and number of steps) we first looked at the combined gesture effect (as opposed to parsing out the effects of instructed vs. spontaneously gesturing) as this is the most powerful analysis to assess the gesture effect hypothesis. In the first step of the stepwise regression analysis we entered age, and VPT (centered) as a predictor, adding gesture prevalence (coding: 0 'no gesture', 1 'gesture') in the second step and the third step the interaction term of the centered VPT and gesture prevalence, as predictors for solving speed. The results of these regression analyses are reported in Table 3. Age, VPT score, gesture prevalence, and the interaction of VPT and gesture prevalence, were unreliable predictors for performance (i.e., solving speed and solving steps) on both the TOH 3-disc and 4-disc.

---

indicate problems for the interpretations of the confirmatory analyses, VIF's for all confirmatory analsyes < 3.05. By common standards, VIF > 10.00 indicate high problems of multicollinearity.

Hypothesis 3. Our third hypothesis was that after controlling for age, instructed and spontaneous gesture prevalence during mental problem-solving would positively affect actual problem-solving performance (solving speed) for all trials (TOH disc 3 and TOH disc 4) as compared to spontaneous non-gesturing, but that this effect would be more pronounced on the most complex task (TOH4) and for participants with a relatively lower visual working memory capacity. We analysed effects of spontaneous vs. no gesture and instructed vs. no gesture prevalence on performance (solving speed and solving steps) on the TOH 3-disc and 4-disc multiple stepwise regression analyses. For each DV (3-disc and 4-disc solving speed and number of steps) we entered age, and VPT (centered) in the first step, and dummy variables spontaneous gesture prevalence (0 = no gesture, 1 = spontaneous gesture) and instructed gesture prevalence (0 = no instructed gesture, 1 = instructed gesture) in the second step. In the third step we entered two interaction terms of the centered VPT with instructed and spontaneous gesture prevalence.

The regression analyses results for hypothesis 3 are shown in Table 4, and each analysis is visualized (without controlling for age) in Figure 3. Again, we did not find any significant results for the overall model fit for this set of predictors on any of the performance measures. Indeed, age, VPT score, spontaneous and instructed gesture prevalence, and the interaction of VPT with spontaneous or instructed gesture prevalence were unreliable predictors for performance (solving speed and solving) on the TOH 3-disc and 4-disc. We did however find one significant interaction effect as predicted; of children scoring lower on the VPT, those who spontaneously gestured needed fewer steps to solve the TOH 4-disc as compared to children scoring lower on the VPT who did not gesture. However, this finding should be interpreted with caution given the low reliability of the overall model fit. In sum, our confirmatory analyses did not replicate previous findings with this younger age sample.

Exploratory Analyses

Effect of instruction versus no instruction on performance

Although we were specifically interested in the effect of spontaneous and instructed gesture problem-solving performance compared to no gesturing, it should be noted that in the current design the spontaneous gesture group and the no gesture group were not randomly assigned by experimental manipulation. Random assignment to those groups would be impossible as our aim was to assess effects of *spontaneous* behavior. One of the reviewers therefore suggested a valuable additional analyses which checks for an effect of the "true" manipulation in the current design. The true manipulation consists of the no instruction versus instructed gesture condition. For this analysis participants were divided into a no instruction condition (i.e. the no gesture group plus the spontaneous gesture group) versus the instructed gesture condition. These analyses could provide insight into whether instruction by itself affected performance. Solving speed of TOH 3-disc for the gesture instruction condition was $M = 33.87$, $SD = 28.77$ versus the no instruction condition, $M = 47.75$, $SD = 47.41$. Solving speed for the TOH 4-disc: $M = 123.76$s, $SD = 69.34$s for the gesture instruction condition and $M = 103.78$s, $SD = 63.59$s for the no instruction condition. Accuracy for the TOH 3-disc was $M = 10.13$, $SD = 5.68$ for the gesture instruction condition and $M = 9.53$, $SD = 4.92$ for the no instruction condition. Accuracy for the TOH 4-disc was $M = 32.23$, $SD = 9.84$ for the gesture instruction condition and $M = 30.46$, $SD = 13.09$ for the no instruction condition.

As to assess whether differences in performance were statistically significant we repeated the confirmatory analyses with this new group variable (no instruction condition versus instructed gesture condition). We did not find any significant differences on performance solving speed or accuracy the TOH3 or the TOH4, before or after controlling for age and visual working

memory capacity, *ps* > .190. To exemplify, the simplest analyses (without covariates) revealed a non-reliable effect of instruction on solving speed on the TOH 3, β = -.158, $t(65) = -1.28$, p = .205, or solving speed on the TOH4, β = .147, $t(57) = 1.113$, p = 270. Full analyses (with covariates) of instruction on performance (accuracy and solving speed) can be obtained at https://osf.io/dreks/ in the folder exploratory analyses.

Strength of evidence for lack of effects of gesture on performance

Null-findings are notoriously difficult to interpret using inferential statistics. Insignificant p-values may simply indicate that the current sample is too small to detect effects of gesture on problem solving performance. Therefore we assessed the strength of the evidence for our null-findings for an effect of gesture, performing an additional Bayesian analyses using JASP, JASP Team (2016), JASP (Version 0.7.5.5).

Gesture (no gesture vs. spontaneous gesture vs. instructed gesture) and its relation with performance on the Tower of Hanoi 3-disc and 4-disc for both speed and steps were reassessed (see Table 3 for results, see also Table 2 for means and standard deviations). ANOVA Bayes factors for null-model (i.e., BF_{01}) are reported with a default priors pre-set by JASP, $p(M) = 0.5$ (Rouder, Morey, Verhagen, Swagman, & Wagenmakers, 2016).

- Insert Table 5 here -

The results suggest that for speed and steps for both TOH 3-disc and TOH 4-disc, the null-hypothesis was favoured over the alternate model predicting an effect of gesture. For speed on the TOH 3-disc, there was substantial evidence for a lack of an effect, Bayes factor BF_{01} = 3.96, which indicates that the observed data are 3.96 more likely under the null-hypothesis as under the alternate model predicting differences between conditions. "Substantial", according to the classification scheme of Jeffrey (1961), which classifies the strength of effects related to

Bayes Factors (no effect [BF = 1], anecdotal evidence [BF = 1-3], substantial evidence [BF = 3-10], strong [BF = 10-30], very strong [BF = 30-100], decisive [BF >100]). Similarly, for steps on TOH 3-disc, there was substantial evidence for an absence of an effect of gesture, Bayes factor $BF_{01} = 7.04$, indicating that the observed data is 7.04 time more likely under the null-hypothesis as under the model that predicts differences in gesture. Similarly, substantial evidence for a lack of an effect of gesture on performance on the Tower of Hanoi 4-disc was also obtained. For speed for TOH 4-disc, there was substantial evidence (Jeffrey, 1961) for the null model, Bayes factor $BF_{01} = 7.04$, which indicates that the observed data are 7.04 more likely under the null-hypothesis as under the alternate model predicting an effect of gesture. Finally, for steps for TOH4, a Bayes factor was obtained of $BF_{01} = 5.607$, indicating that the observed data are 5.607 more likely under the null-hypotheses. In sum, there is evidence that gesture did not affect problem-solving performance in the current sample.

Further note that independent Bayesian t-tests of the simpler contrast of gesture (spontaneous or instructed) versus no gesture on performance on the TOH 3-disc and TOH 4-disc reveal similar substantial evidence that the data were more likely under the null-model predicting that no effect was present as compared to the alternative model predicting a positive effect of gesture on performance (i.e., null-hypothesis, TOH 3 speed $BF_{01} = 6.80$, TOH 3 steps $BF_{01} = 4.75$, TOH 4 speed $BF_{01} = 3.675$, TOH 4 steps $BF_{01} = 4.252$, Bayes factors based on a Cauchy prior of $h = .75$ for the alternate model (default prior set by JASP); analyses can be retrieved from https://osf.io/dreks/). Finally note, that our interpretation of the data are relatively robust at using priors that assume a lower effect size of gesturing (see Figure 4 for more information).

- Insert Figure 4 here –

Discussion

In the current study we aimed to replicate previous findings with adults (Pouw et al., under review) in young children. In adults, participants with lower visual working memory capacity who gestured either spontaneously or when instructed to do so during mental problem solving, performed better when subsequently solving the more difficult Tower of Hanoi (TOH) problem than participants who did not gesture. In the current age sample with 8-12 year old children, we did not replicate these findings. That is, there were no effects of gesture (either spontaneous or instructed) on problem solving performance, and there was substantial evidence for the absence of an effect. As such it is unsupported that gesture affected problem solving performance, independently, or as a function of visual working memory capacity and task difficulty.

There are several possible reasons why our hypothesized gesture effect was not replicated in the current sample. Firstly, in contrast to the study with adults, children's problem-solving performance was not correlated with performance on the visual working memory task. One possible explanation is that we used a more simplified version of the task, wherein children recreated the visual pattern by selecting locations through mouse-clicking, whereas the study with adults required verbally recalling the pattern by naming the letters that were assigned to the previous locations during the response phase. This difference in the task potentially recruits different cognitive processes. Given that we assume that gestures become effective when task-relevant resources are taxed, if we failed to gauge such task-relevant resources with the current task, then the current study was not able to test the cognitive load hypothesis. However, we doubt whether the difference of the simplified task with the task used with adults recruits different resources in such a dramatic way. Yet, problematically, this does beg the question why

children's visual working memory resources were not correlated with task performance. One possible explanation is that children may use a different strategy than adults, and use more various strategies that do not recruit such visual imagery processes. Indeed it has been argued that children of the current age group are still developing the planning skills that are required to solve the Tower of Hanoi (Schiff & Vakil, 2015).

It should be noted that we found statistically significant relationships between the number of gestures that children produced and performance on the TOH 3-disc problem (solving time & solving steps), but only when children were instructed to gesture. However, spontaneous gesture rates were not correlated with performance. These mixed findings (together with a lack of an overall performance effect of gesture) caution us to interpret the positive correlation with instructed gesture rate with performance as an effect of gesture per se. It could for example be that children that are more likely to keep following up instructions (gesturing more) are also more poised to perform better. This is currently an open question that could have been excluded would we have observed an effect of spontaneous gesturing on performance. Nevertheless, it is telling that the correlation of instructed gesture rate and performance disappears for the more difficult task. It is clear that children were having difficulty with the current TOH 4-disc problem as compared to the TOH 3-disc, with 62% of the children solving the TOH 3-disc in the minimal number of moves as compared to 5% in the TOH 4-disc problem. This could indicate that as the task becomes too difficult, that the effect of gesture dissipates. Chu & Kita (2011) have argued that a possible explanation for why gestures are not affected during more difficult problems is because the problem solver must use extra cognitive resources to recruit gestures. As such, the current results can support the idea that such cognitive resources for recruiting gestures can be

productive in less (i.e., more optimally) challenging problems (e.g., TOH 3-disc), when the task becomes too difficult such cognitive resources are too costly to spend.

A possible limitation of our study is that the participants in the No Gesture group were not instructed to refrain from gesturing. In other words, non-gesturing participants were not randomly "assigned" to their condition. The reason for this approach is that we wanted to rule out that any positive effect of gesturing would not be due to gesturing being helpful, but rather to inhibiting gesturing being harmful for performance (Hostetter & Alibali, 2008). One might argue that it could be that children who have more difficulties with the task are more likely to gesture, which might obscure a true effect of gesture as they are already poised to perform worse on the task. However, even if this is the case, since we manipulated gesture use, we still should have found an increased performance overall due to gesturing when instructed, relative to the non-instructed gesture group. This was not the case.

A more general worry of the present results is that a younger age sample inevitably produces more noisy data, which might lower the detection of a possible gesture effect. One way that this manifested itself is that some participants seemed more engaged with the task than others (i.e., some children were not very enthused about performing the mental preparation task for [the full] 150 seconds). Even though we did not obtain any differences in self-reported motivation scores across conditions, given that we are unable to assess with certainty how engaged children are during the mental preparation phase, it might be that this natural variability prevents detection of a potential gesture effect. Yet, in any case the current sample does provide information about the ecological robustness of a potential co-thought gesture effect in children; which would not be very robust in a sample comparable to the present one (8-12 yrs old).

Although the results did not confirm our hypotheses, the current study provides novel evidence that children (in comparable ways to adults) spontaneously adopt co-thought gestures when being confronted with mentally simulating the problem space of the Tower of Hanoi. This is interesting, as it provides evidence that co-thought gesturing is already part of the cognitive toolkit in earlier development, and such gesturing persists throughout adulthood (Pouw et al., under review). Thus, at a minimum this study provides a productive paradigm that naturally solicits co-thought gestures from children, which will be useful for the further investigation of the cognitive role of co-thought gesticulation in younger age samples Furthermore, it is interesting that similar to the adult sample (Pouw et al., under review), children were spontaneously producing pointing (deictic) gestures. This contrasts with findings on co-speech gestures where participants explain their problem solving. The observed gestures in these studies are predominately iconic in nature; gestures simulating actions on an imagined Tower of Hanoi (e.g., Beilock & Goldin-Meadow, 2010; Trofatter, Kontra, Beilock, & Goldin-Meadow, 2015). Such findings on co-speech gestures are not directly comparable however. Firstly, they cannot be compared on effects of performance as these co-speech gestures are held to affect performance when dealing with manipulative changes in the TOH apparatus (changes in disc size; see Beilock & Goldin-Meadow, 2010). In the current study we did not make changes in manipulability of the task. Additionally, a key difference with studies on co-speech gestures is that in the current task the TOH apparatus was not present during gesturing, while in the current study the TOH apparatus was available. This availability of the TOH apparatus is likely to explain why pointing gestures are observed in the current study together with the fact that the current gestures are not communicative in nature.

There is another aspect that needs to be emphasized to put the current results in an appropriate context. Firstly, in the current study we did not compare the effect of gesture with the inhibition of gesture, but rather with spontaneous non-gesturing. We reasoned that if we did obtain an effect of gesture it would allow us to conclude that the production rather than the inhibition of gestures affects cognitive processing (for a discussion of the theoretical importance of this difference, see Cook, Yip, & Goldin-Meadow, 2010; Goldin-Meadow et al., 2001; Pouw et al., 2014; Wagner, Nusbaum, & Goldin-Meadow, 2004). However, future studies could compare the role of inhibiting gestures as well, to see whether gesture effects on performance do arise. That is, given the fact that children did spontaneously gesture in the current study, it could be the case that actively inhibiting children to move their hands would result in a gesture effect on performance (albeit in a negative way).

Conclusion

The current study does provide more insight on the role of co-thought gestures in thinking and the potential boundary conditions concerning the beneficial effect on problem solving. Namely, children (8-12 yrs) spontaneously use gestures in similar ways (pointing gestures), and in similar amounts (Figure 2) as compared to adults, although in contrast to adults, these gestures do not positively affect children's problem solving as compared to not gesturing. As such, it keeps the question that motivated this study in the first place very alive: what is the cognitive function of co-thought gesture? Why does the current task invoke spontaneous pointing gestures, while others evoke more iconic pantomimic gestures (e.g., Chu & Kita, 2008)? Moreover, the current study can inspire a more systematic study into the development of co-thought gestures in children, as it provides a paradigm in which these gestures are naturally adopted. This is needed, as current studies on the developmental emergence/ontogenetics of

gestures largely ignore the phenomenon of co-thought gestures (e.g., Goldin-Meadow, 1998; Kendon, 2004; McNeill, 2008).

References

Beilock, S. L., & Goldin-Meadow, S. (2010). Gesture changes thought by grounding it in action. *Psychological Science*, *21*(11), 1605–1610.

Chu, M., & Kita, S. (2008). Spontaneous gestures during mental rotation tasks: Insights into the microdevelopment of the motor strategy. *Journal of Experimental Psychology: General*, *137*(4), 706-723.

Chu, M., & Kita, S. (2011). The nature of gestures' beneficial role in spatial problem solving. *Journal of Experimental Psychology: General*, *140*(1), 102-115.

Chu, M., & Kita, S. (2015). Co-thought and co-speech gestures are generated by the same action generation process. Advance online publication.

Chu, M., Meyer, A., Foulkes, L., & Kita, S. (2014). Individual differences in frequency and saliency of speech-accompanying gestures: the role of cognitive abilities and empathy. *Journal of Experimental Psychology: General*, *143*(2), 694.

Cook, S. W., Yip, T. K., & Goldin-Meadow, S. (2012). Gestures, but not meaningless movements, lighten working memory load when explaining math. *Language and Cognitive Processes*, *27*(4), 594–610.

Delgado, B., Gómez, J. C., & Sarriá, E. (2011). Pointing gestures as a cognitive tool in young children: experimental evidence. *Journal of Experimental Child Psychology*, 110, 299–312.

Della Sala, S., Gray, C., Baddeley, A., & Wilson, L. (1997). *The Visual Patterns Test: A new test of short-term visual recall*. Bury St. Edmunds, England: Thames Valley Test.

Gathercole, S. E., Pickering, S. J., Ambridge, B., & Wearing, H. (2004). The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, *40*(2), 177-190.

Goldin-Meadow, S. (1998). The development of gesture and speech as an integrated system. *New Directions for Child and Adolescent Development, 79*, 29-42.

Goldin-Meadow, S., Nusbaum, H., Kelly, S. D., & Wagner, S. (2001). Explaining math: Gesturing lightens the load. *Psychological Science*, *12*(6), 516–522.

Hegarty, M., Mayer, S., Kriz, S., & Keehner, M. (2005). The role of gestures in mental animation. *Spatial Cognition and Computation, 5*, 333–356.

Hostetter, A. B., & Alibali, M. W. (2008). Visible embodiment: Gestures as simulated action. *Psychonomic Bulletin & Review*, *15*(3), 495–514.

Hostetter, A. B., Alibali, M. W., & Kita, S. (2007). I see it in my hands' eye: Representational gestures reflect conceptual demands. *Language and Cognitive Processes, 22*, 313–336.

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

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge University Press.

Kita, S., Alibali, M. W., & Chu, M. (in press). How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological Review*.

Logan, T., Lowrie, T., & Diezmann, C. M. (2014). Co-thought gestures: Supporting students to successfully navigate map tasks. *Educational Studies in Mathematics, 87*, 87-102.

Marstaller, L., & Burianová, H. (2013). Individual differences in the gesture effect on working memory. *Psychonomic Bulletin & Review*, *20*, 496–500.

McNeill, D. (2008). *Gesture and thought*. University of Chicago Press.

Paas, F., & Sweller, J. (2012). An evolutionary upgrade of cognitive load theory: Using the human motor system and collaboration to support the learning of complex cognitive tasks. *Educational Psychology Review*, *24*(1), 27-45.

Paas, F., Tuovinen, J. E., Tabbers, H., & Van Gerven, P. W. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist*, *38*(1), 63-71.

Pouw, W. T. J. L., De Nooijer, J. A., Van Gog, T., Zwaan, R. A., & Paas, F. (2014). Toward a more embedded/extended perspective on the cognitive function of gestures. *Frontiers in Psychology*, *5*, 359.

Pouw, W. T. J. L., Eielts, C., van Gog, T., Zwaan, R. A., & Paas, F. (under review). Problem solving is supported by co-thought gesticulation when task complexity is high and visual (but not spatial) working memory capacity is low.

Pouw, W. T. J. L., Mavilidi, M., Van Gog, T., & Paas, F. (2016). Gesturing during mental problem solving reduces eye movements, especially for individuals with lower visual working memory capacity. *Cognitive Processing*, *17*(3), 269-277.

Rouder, J. N., Morey, R. D., Verhagen, J., Swagman, A. R., & Wagenmakers, E. J. (2016). Bayesian analysis of factorial designs. *Psychological Methods*. Advance online publication. doi: 10.1037/met0000057

Schiff, R., & Vakil, E. (2015). Age differences in cognitive skill learning, retention and transfer: The case of the Tower of Hanoi Puzzle. *Learning and Individual Differences*, *39*, 164-171.

Schwartz, D. L., & Black, T. (1999). Inferences through imagined actions: Knowing by simulated doing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 116–136.

So, W. C., Shum, P. L. C., & Wong, M. K. Y. (2015). Gesture is more effective than spatial language in encoding spatial information. *The Quarterly Journal of Experimental Psychology*, *68*(12), 2384-2401.

Trofatter, C., Kontra, C., Beilock, S., & Goldin-Meadow, S. (2015). Gesturing has a larger impact on problem-solving than action, even when action is accompanied by words. *Language, Cognition and Neuroscience*, *30*(3), 251–260.

Wagner, S. M., Nusbaum, H., & Goldin-Meadow, S. (2004). Probing the mental representation of gesture: Is handwaving spatial? *Journal of Memory and Language*, *50*(4), 395-407.

Table 1. *Means (and SD) and Correlations of Problem-Solving Performance, VPT, and Gesture*.

Overall	Mean(*SD*)	1.	2.	3.	4.	5.
1. VPT	.41(.20)		-.074	-.014	-.328	**-.467***
2. TOH 3-disc Solving Time	47.75(47.42)			**.734****	.376	.244
3. TOH 3-disc Solving Steps	9.95(4.92)				.238	.127
4. TOH 4-disc Solving Time	103.78(63.59)					.127
5. TOH 4-Solving Steps	30.46(13.09)					
No Gesture Group TOH 3-disc (*N* = 29)	Mean(*SD*)	1.	2.	3.		
1.VPT	.41(.20)		-.174	-.004		
2. TOH 3-disc Solving Time	48.63(46.91)			**.715****		
3. TOH 3-disc Solving Steps	9.97(5.28)					
Spontaneous Gesture Group TOH 3-disc (*N* = 13)	Mean(*SD*)	1.	2.	3.	4.	
1. VPT	.41(.20)		-.361	-.360	.106	
2. TOH 3-disc Solving Time	48.15(51.52)			**.880***	-.042	
3. TOH 3-disc Solving Steps	10.15(4.36)				-.154	
4. TOH 3-disc Gesture Frequency	48.85(48.92)					
Instructed Gesture Group TOH 3-disc (*N* = 23)	Mean(*SD*)	1.	2.	3.	4.	
1. VPT	.45(.21)		-0.74	-.014	.152	
2. TOH 3-disc Solving Time	33.87(28.77)			**.734****	**-.634****	
3. TOH 3-disc Solving Steps	9.44(4.51)				**-.483***	
4. TOH 3-disc Gesture Frequency	77.13(29.37)					
No Gesture Group TOH 4-disc (*N* = 21)	Mean(*SD*)	1.	2.	3.		
1.VPT	.42(.21)		**-.469***	-.404		
2. TOH 4-disc Solving Time	111.24(63.78)			**.95****		
3. TOH 4-disc Solving Steps	31.57(13.59)					
Spontaneous Gesture Group TOH 4-disc (*N* = 16)	Mean(*SD*)	1.	2.	3.	4.	
1. VPT	.45(.15)		-.070	.347	.105	
2. TOH 4-disc Solving Time	94.00(64.05)			**.878****	-.333	
3. TOH 4-disc Solving Steps	29.00(12.70)				-.258	
4. TOH 4-disc Gesture Frequency	44.19(37.42)					
Instructed Gesture Group TOH 4-disc (*N* = 21)	Mean(*SD*)	1.	2.	3.	4.	
1. VPT	.45(.21)		-.328	**-.467***	.231	
2. TOH 4-disc Solving Time	123.76(69.34)			.738	-.243	
3. TOH 4-disc Solving Steps	32.23(9.84)				-.148	
4. TOH 4-disc Gesture Frequency	73.381(31.73)					

Note. ** $p < .01$, * $p < .05$. When TOH 4-disc is concerned we only report means and correlations for participants who were able to solve the task.

Table 2. *Means (and SD) of Self-reported Difficulty, Mental Effort, and Interest*

Overall	TOH 3-disc Mean(*SD*)	TOH 4 Hanoi 4-disc Mean(*SD*)
1. Difficulty	2.56(0.93)	3.86(0.90)
2. Mental Effort	3.02(1.09)	4.08(0.89)
3. Interest	4.21(0.81)	4.42(0.70)
No Gesture Group	Mean(*SD*)	Mean(*SD*)
1. Difficulty	2.41(0.63)	3.75(0.99)
2. Mental Effort	2.76(1.09)	3.92(1.02)
3. Interest	4.240.83)	4.38(0.71)
Spontaneous Gesture Group	Mean(*SD*)	Mean(*SD*)
1. Difficulty	2.77(1.09)	3.89(0.88)
2. Mental Effort	3.46(0.78)	4.11(0.81)
3. Interest	4.08(0.95)	4.37(0.68)
Instructed Gesture Group	Mean(*SD*)	Mean(*SD*)
1. Difficulty	2.57(1.12)	3.95(0.84)
2. Mental Effort	3.13(1.18)	4.23(0.81)
3. Interest	4.21(0.74)	4.50(0.74)

Table 3. *Analyses of Hypothesis 1 & 2*

Solving Speed TOH 3-disc	Step	B	(*SE*)	β	*t* (*df = 62, 61, 60)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	109.376	51.580		2.121	**.038**	.034	.127
2. Age		-6.696	5.215	-.188	-1.284	.204		
3. VPT (Centralised)		-20.511	31.019	-.097	-.661	.511		
1. Constant	2	107.996	51.867		2.082	**.042**	.025	.211
2. Age		-6.170	5.302	-.173	-1.164	.249		
3. VPT (Centralised)		-21.294	31.188	-.100	-.683	.497		
4. Gesture (0 = no gesture, 1 = gesture)		-6.854	10.576	-.081	-.648	.519		
1. Constant	3	108.005	52.299		2.065	**.043**	.009	.346
2. Age		-6.171	5.346	-.173	-1.154	.253		
3. VPT (Centralised)		-20.690	43.135	-.098	-.480	.633		
4. Gesture (0 = no gesture, 1 = gesture)		-6.852	10.664	-.081	-.642	.523		
5. Interaction VPT & Gesture		-1.089	53.172	-.004	-.020	.984		
Solving Steps TOH 3-disc	Step	B	(*SE*)	β	*t* (*df = 62, 61, 60)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	6.312	5.979		1.056	.295	-.021	.713
2. Age		.357	.604	.089	.590	.557		
3. VPT (Centralised)		-2.892	3.596	-.121	-.804	.424		
1. Constant	2	6.241	6.029		1.035	.305	-.036	.861
2. Age		.384	.616	.095	.623	.536		
3. VPT (Centralised		-2.932	3.625	-.123	-.809	.422		
4. Gesture (0 = no gesture, 1 = gesture)		-.352	1.229	-.037	-.286	.776		
1. Constant	3	6.265	6.068		1.032	.306	-.050	.914
2. Age		.382	.620	.095	.617	.540		
3. VPT (Centralised)		-1.324	5.005	-.055	-.264	.792		
4. Gesture (0 = no gesture, 1 = gesture)		-.346	1.237	-.036	-.280	.781		
5. Interaction VPT & Gesture		-2.897	6.169	-.090	-.470	.640		
Solving Speed TOH 4-disc	Step	B	(*SE*)	β	*t* (*df =55, 54, 53)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	187.002	80.267		2.330	**.024**	.083	**.034**
2. Age		-7.475	8.106	-.140	-.922	.360		
3. VPT (Centralised)		-81.942	51.162	-.243	-1.602	.115		
1. Constant	2	185.963	83.492		2.227	.030	.066	.083

2. Age		-7.428	8.232	-.139	-.902	.371		
3. VPT (Centralised)		-82.262	52.006	-.243	-1.582	.120		
4. Gesture (0 = no gesture, 1 = gesture)		.901	17.536	.007	.051	.959		
1. Constant	3	182.121	84.271		2.161	**.035**	.055	.138
2. Age		-7.031	8.311	-.131	-.846	.401		
3. VPT (Centralised)		-114.036	75.769	-.338	-1.505	.138		
4. Gesture (0 = no gesture, 1 = gesture)		.194	17.687	.001	.011	.991		
5. Interaction VPT & Gesture		51.580	88.955	.118	.580	.564		
Solving Steps TOH 4-disc	Step	B	(*SE*)	β	*t* (*df = 55, 54, 53)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	39.973	14.973		2.670	**.010**	.033	.150
2. Age		-.867	1.512	-.089	-.573	.569		
3. VPT (Centralised)		-12.164	9.544	-.198	-1.275	.208		
1. Constant	2	40.585	15.571		2.606	**.012**	.015	.286
2. Age		-.895	1.535	-.092	-.583	.562		
3. VPT (Centralised)		-11.976	9.699	-.195	-1.235	.222		
4. Gesture (0 = no gesture, 1 = gesture)		-.531	3.270	-.022	-.162	.872		
1. Constant	3	39.247	15.592		2.517	**.015**	.019	.293
2. Age		-.757	1.538	-.078	-.492	.625		
3. VPT (Centralised)		-23.036	14.019	-.375	-1.643	.106		
4. Gesture (0 = no gesture, 1 = gesture)		-.777	3.272	-.031	-.237	.813		
5. Interaction VPT & Gesture		17.955	16.459	.225	1.091	.280		

Table 4. *Analyses of Hypotheses 1 & 3*

Solving Speed TOH 3-disc (see Figure 3 a)	Step	B	(*SE*)	*β*	*t* (*df = 62,60, 58)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	109.376	51.580		2.121	**.038**	.064	.127
2. Age		-6.696	5.215	-.188	-1.284	.204		
3. VPT (Centralised)		-20.511	31.019	-.097	-.661	.511		
1. Constant	2	113.119	52.113		2.171	**.034**	.086	.241
2. Age		-6.699	5.327	-.188	-1.257	.213		
3. VPT (Centralised)		-17.888	31.370	-.084	-.570	.571		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		2.649	14.190	.025	.187	.853		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		-12.030	11.764	-.137	-1.023	.311		
1. Constant	3	117.751	52.494		2.243	**.029**	.108	.332
2. Age		-7.178	5.367	-.201	-1.338	.186		
3. VPT (Centralised)		-17.506	43.023	-.083	-.407	.686		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		2.317	14.266	.022	.162	.872		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		-12.757	11.840	-.145	-1.077	.286		
6. Interaction VPT & Spontaneous Gesture		-57.844	72.731	-.118	-.795	.430		
5. Interaction VPT & Instructed Gesture		32.810	58.869	.094	.557	.579		
Solving Steps TOH 3-disc (see Figure 3 b)	Step	B	(*SE*)	*β*	*t* (*df = 62, 60, 58)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	6.312	5.979		1.056	.295	-.021	.713
2. Age		.357	.604	.089	.590	.557		
3. VPT (Centralised)		-2.892	3.596	-.121	-.804	.424		
1. Constant	2	6.437	6.103		1.055	.296	-.052	.913
2. Age		.363	.624	.090	.583	.562		
3. VPT (Centralised)		-2.802	3.674	-.117	-.763	.449		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		.012	1.662	.001	.007	.994		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		-.549	1.378	-.055	-.399	.691		
1. Constant	3	6.813	6.178		1.103	.275	-.072	.943
2. Age		.326	.632	.081	.516	.608		
3. VPT (Centralised)		-1.145	5.063	-.048	-.226	.822		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		-.040	1.679	-.003	-.024	.981		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		-.585	1.393	-.059	-.420	.676		
6. Interaction VPT & Spontaneous Gesture		-7.558	8.560	-.136	-.883	.381		
5. Interaction VPT & Instructed Gesture		-.335	6.928	-.008	-.048	.962		

Solving Speed TOH 4-disc (see Figure 3c)	Step	B	(*SE*)	β	*t* (*df = 55, 53, 51)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	187.002	80.267		2.330	**.024**	.083	**.034**
2. Age		-7.475	8.106	-.140	-.922	.360		
3. VPT (Centralised)		-81.942	51.162	-.243	-1.602	.115		
1. Constant	2	204.768	83.195		2.461	**.017**	.092	.058
2. Age		-9.308	8.205	-.174	-1.134	.262		
3. VPT (Centralised)		-75.900	51.448	-.225	-1.475	.146		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		-18.481	21.179	-.127	-.873	.387		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		14.885	19.414	.110	.767	.447		
1. Constant	3	197.225	84.905		2.323	**.024**	.067	.145
2. Age		-8.542	8.377	-.160	-1.020	.313		
3. VPT (Centralised)		-107.979	75.492	-.320	-1.430	.159		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		-20.377	21.640	-.140	-.942	.351		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		14.755	19.760	.109	.747	.459		
6. Interaction VPT & Spontaneous Gesture		97.373	127.453	.118	.764	.448		
5. Interaction VPT & Instructed Gesture		31.584	94.911	.061	.333	.741		
Solving Steps TOH 4-disc (see Figure 3d)	Step	B	(*SE*)	β	*t* (*df = 55, 53, 51)*	*p*	$R^2_{adjusted}$	*p* Model
1. Constant	1	39.973	14.973		2.670	**.010**	.033	.150
2. Age		-.867	1.512	-.089	-.573	.569		
3. VPT (Centralised)		-12.164	9.544	-.198	-1.275	.208		
1. Constant	2	42.647	15.755		2.707	**.009**	.012	.331
2. Age		-1.101	1.554	-.113	-.709	.482		
3. VPT (Centralised)		-11.278	9.743	-.184	-1.158	.252		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		-2.656	4.011	-.100	-.662	.511		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		1.003	3.676	.041	.273	.786		
1. Constant	3	38.478	15.306		2.514	**.015**	.08	.112
2. Age		-.680	1.510	-.070	-.450	.654		
3. VPT (Centralised)		-23.345	13.609	-.380	-1.715	.092		
4. Spontaneous Gesture (0 = no gesture, 1 = spontaneous gesture)		-3.850	3.901	-.145	-.987	.328		
5. Instructed Gesture (0 = no gesture, 1 = instructed gesture)		1.201	3.562	.049	.337	.737		
6. Interaction VPT & Spontaneous Gesture		53.566	22.976	.358	2.331	**.024**		
5. Interaction VPT & Instructed Gesture		4.247	17.110	.045	.248	.805		

Table 5. Bayesian ANOVA.

Model Comparison - TOH3speed					
Models	P(M)	P(M|data)	BF_M	BF_{01}	% error
Null model	0.500	0.798	3.957	1.000	
Gesture	0.500	0.202	0.253	3.957	0.032
Model Comparison - TOH3steps					
Models	P(M)	P(M|data)	BF_M	BF_{01}	% error
Null model	0.500	0.876	7.039	1.000	
Gesture	0.500	0.124	0.142	7.039	0.027
Model Comparison - TOH4speed					
Models	P(M)	P(M|data)	BF_M	BF_{01}	% error
Null model	0.500	0.787	3.695	1.000	
Gesture	0.500	0.213	0.271	3.695	0.035
Model Comparison - TOH4steps					
Models	P(M)	P(M|data)	BF_M	BF_{01}	% error
Null model	0.500	0.849	5.607	1.000	
Gesture	0.500	0.151	0.178	5.607	0.030

Figure titles and legends

Figure 1. Two examples of spontaneous gestures arising during mental problem solving for the TOH (1 frame per second). Faces blurred for anonymisation.

Figure 2. Gesture frequency, and spontaneous gesture likelihood on the TOH 3-disc and 4-disc. Gesture production data for the current sample with children versus the adult sample as reported in (Pouw et al., under review). Next to the gesture rate during each mental preparation trial (150 seconds of mental preparation) per age group, the number of spontaneous gesturers (i.e., the number of participants from the total no instruction group that spontaneously adopted gesture) is presented.

Figure 3a-d. Confirmatory analysis graphs.

Figure 4. T-test Bayes factors for gesture versus no gesture effect on TOH3 (left) and TOH4 (right) performance (solving speed). The figures show the robustness of the Bayes Factor for the two separate T-tests as a function of setting the Cauchy prior width. Lower (higher) widths indicate higher certainty (uncertainty) of the effect size (centered around zero) assuming the alternative hypothesis (presence of an effect) is true. The grey dots indicate the reported Bayes factor at the default Cauchy prior width of .71. Figures were produced by JASP, jasp-stat.org.