Differential influences of visuospatial and phonological resources on mental arithmetic

Edward H. Chen, Susanne M. Jaeggi, & Drew H. Bailey

University of California, Irvine

Abstract

Separate but related cognitive processes are hypothesized to influence different arithmetic operations. Working memory has been compartmentalized into a number of different sub-processes, such as phonological memory and visuospatial memory that are believed to have unique contributions to the performance of two distinct arithmetic operations: multiplication and subtraction. Influential work had initially produced these effects, but subsequent experiments have yielded inconsistent results. Because the reasons for these inconsistencies are not immediately apparent, the current study aims to systematically review these subsequent attempts and propose an experimental design to investigate a differential effect between working memory processes and arithmetic operation. This report will attempt to replicate this effect using tasks developed from prior work and perform analyses across a number of different subsamples. Details for predictions, methods, and analytic plans are described further below.

Key words: working memory, math cognition, dual task, arithmetic

Differential influences of visuospatial and phonological resources on mental arithmetic

Evidence from cognitive psychology and neuroscience suggests domain-specific components of working memory may contribute to differences in mental arithmetic performance, but several important questions remain unanswered. A number of imaging and lesion studies suggest the parietal regions are heavily involved with the process of mental arithmetic, specifically addition and subtraction as well as with visuospatial processes (Dehaene & Cohen, 1997; Dehaene et al., 2003; Prado et al., 2011). Meanwhile, additional evidence suggests that another arithmetic operation, namely multiplication, relies on different neural substrates found within the perisylvian areas which have been found to modulate phonological and language processes (Dehaene & Cohen, 1997; Dehaene et al., 2003; Kawashima et al., 2004, Prado et al., 2011). These would suggest that visuospatial processes are involved with subtraction while phonological resources are involved in multiplication; however, behavioral experiments do not paint this exact picture.

The current study will review these approaches and their findings and describe our current approach to investigate the unique contributions of working memory within mental arithmetic. An influential study by Lee & Kang (2002) investigated a differential effect of working memory resources on arithmetic operation type. In their study, participants were given single-digit multiplication and subtraction trials where answers were typed in using a number pad. Participants performed arithmetic in three conditions: with no secondary task, while repeating a non-word string (phonological (PL) load), or while remembering the shape and position of an object (visuospatial (VSSP) load). They reported very large effects indicating that Korean undergraduates' multiplication performance was worse than subtraction under

phonological load (Cohen's d = 2.39[1]; Table 1). A similarly large but opposite effect was reported where subtraction performance was worse than multiplication under visuospatial load (d = 3.35). Interestingly, the effect of PL load on subtraction relative to subtraction alone was almost 0, as was the effect of VSSP load on multiplication relative to multiplication alone.. They predicted arithmetic operations to be facilitated through specific modular representations; that is, multiplication is enacted through an auditory-phonological encoding while subtraction is enacted through an analog magnitude system like a mental number line. This line of reasoning is consistent with parallel processing theories of dual-task performance which ascribe differences in reaction time and accuracy performance to domain-specific resources competing for space within the working memory (Navon & Miller, 1987; Pashler, 1994). In other words, the more similar two tasks appear with regards to the overlap between the demands of the primary task and the modality imposed demands of the secondary task, such as a visuospatial span task with a visual imagery task, the more interference we should observe.

Several studies have used similar methods; although some have replicated the direction of these effects, none have produced the pattern of opposite effects with magnitudes approaching the size in the original Lee & Kang (2002) study. Strikingly, while there is variation in the kinds of tasks and samples among others, the original study was not especially different in its design that would lead to the discrepancy in effect sizes (Table 1). Neither of the two partial replications used an entirely within-subject design like Lee & Kang (2002), which could have potentially led to the discrepancy in effect sizes. The current paper will improve upon Lee & Kang (2002) and

[1] Cohen's *d* was calculated by hand. RT means were taken from reported values within Lee & Kang while SD were calculated from reported standard errors (SD = SE*$\sqrt{n}$). Thus, we used the following values: multiplication under phonological load (M = 1169.5, SD = 82.85); subtraction under phonological load (M = 993, SD = 63.56). Values were then input into the classic Cohen's *d* formula: d = $\frac{M_1 - M_2}{\sigma}$, where σ is the pooled standard deviation of the two means: pooled SD = $\sqrt{\frac{SD_1^2 + SD_2^2}{2}}$. The same method was used to calculate the effect size for visuospatial load.

the previous replication attempts by using an entirely within-subjects design and by using a larger sample size than any of the previous studies. Imbo & LeFevre (2010) attempted to replicate the findings using a mix of native Chinese and Canadian participants to perform arithmetic problems under load (see Table 1 for details). They found differential impacts of phonological and visuospatial loads in Chinese students attending a Canadian university but not in other Canadian students. However, the interaction was only found in multiplication errors such that multiplication was less accurate in the Chinese students compared to Canadian students when under phonological cognitive load. While the effect of visuospatial load was not found in subtraction, Chinese students exhibited decreased performance compared to Canadian students on the secondary visuospatial task when arithmetic was presented vertically. While multiplication was affected by PL load, subtraction should have been impaired by VSSP load due to students having abacus training. Differences in performance were attributed to cultural differences in education, such as the use of the rhyming song many Chinese students use to learn multiplication which requires phonological resources (Imbo & LeFevre, 2010, p. 183). Meanwhile, authors hypothesized that learning addition and subtraction on an abacus, a more common practice in China than Canada, causes students to use strategies that require greater visuospatial resources.

Considering the variation in design features, the inconsistent results from previous attempts to replicate Lee & Kang (2002) have been attributed to a number of possible reasons. First, a lack of balancing the cognitive demands of the working memory and arithmetic tasks within and across participants raises uncertainty over whether it was the difficulty or specific modality of secondary tasks that led to the interaction reported in Lee & Kang (2002). The use of different multiplication and subtraction tasks as well as WM tasks mask the extent to which

modality effects are separate from the inherent demands of the tasks themselves. Cavdaroglu & Knops (2016) attempted to resolve this issue by having German participants perform arithmetic under similar load conditions to Lee & Kang (2002). Importantly, they created two difficulty conditions that were individually determined through psychometric functions to ensure participants were performing symmetrically difficult secondary tasks. In addition, their calculation tasks attempted to minimize central executive resources by controlling for problem size and difficulty. Under these conditions, their results yielded no differential impact of working memory resources on multiplication and subtraction. Despite claims that the most prominent dissociations exist between multiplication and subtraction (see Lee & Kang, 2002; Lee 2000), these results suggest the validity of the domain-specific working memory influences on mental arithmetic is not as clear. However, difficulty alone may not fully explain the disparity in effect sizes. These previous replication attempts have used different working memory tasks to load the PL and VSSP, so it is not clear either whether the tasks in Lee & Kang (2002) happened to load the WM components more than the replication attempts. Third, the original study included Korean participants, whose math education differs from U.S. and Canadian samples. As evident from Imbo & LeFevre (2010), the Chinese participants who share similarities with Koreans in number system and arithmetic education (e.g., favoring rote memorization through drilling and songs and some mental-abacus training) were the only population that saw a selective interaction effect while the Canadian participants did not. The automaticity gained through extensive practice using specific representational strategies (i.e. phonologically-based rhyming songs and visuospatially-based mental-abacus) in Chinese students was believed to cause a stronger connection between arithmetic operations and specific working memory components. In comparison, Imbo & LeFevre (2010) argued that western math education caused students to use

more variable strategies suggesting a weaker link between specific components and arithmetic but a stronger link to executive resources.

Moreover, current meta-analytic evidence of dual-task experiments also suggest that the influence of specific working memory components on arithmetic performance may not be as robust as other findings related to dual-task performance, such as the effect of domain-general demands of the secondary task on performance (Chen & Bailey, 2021). Specifically, it appears that larger effect sizes between different combinations of WM load and arithmetic may be partly driven by researchers predicting larger effects for more demanding secondary tasks (e.g., those that require more central executive processing). In part because of the large number of researcher degrees of freedom in selecting potential interactions in dual task arithmetic experiments, the robustness of these findings warrants further scrutiny. In summary, it is unclear whether the results from replication attempts reflected important insights regarding arithmetic cognition or if they reflected idiosyncratic aspects of Lee & Kang's (2002) study, specific to a combination of the tasks and sample. Thus, it is imperative to establish better practice towards registering planned analyses in the future.

Current Study

While Cavdaroglu & Knops (2016) improved upon the original design of Lee & Kang (2002), some remaining issues need to be experimentally investigated. The current design will improve upon Cavdaroglu & Knops (2016) in a number of ways. First, the arithmetic condition will be a within- rather than between-subject factor design. It should be noted that this is only fully true when there are no differential sequence effects to be expected, thus we have carefully randomized and counterbalanced the order of conditions and will perform additional analyses to follow-up our main analysis. Specifically, we will test for the key interaction (i.e. load type ×

arithmetic operation) for the first arithmetic under load condition within each participant. Second, it was unclear from Imbo & LeFevre (2010) whether cultural differences in arithmetic performance were confounded by the particular tasks used, so this study will re-examine cultural differences in arithmetic cognition by recruiting students who received their primary math education in China as well as participants who received their primary math education in the U.S.

In the current study, a dual-task paradigm will be used to test the involvement of phonological and visuospatial resources within mental subtraction and multiplication. The aim of this study is to test whether the findings reported in Lee & Kang (2002) can be replicated using similar procedures and tasks as Cavdaroglu & Knops (2016). Participants will solve either multiplication or subtraction problems under phonological (i.e. remembering a string of letters or numbers) and visuospatial load (i.e. remembering the positions of dots in an array). The interaction between these memory load types and operation types was most prominent in Lee & Kang (2002). However, attempts to replicate this large dissociation since have not been wholly successful (see Table 1). Task difficulty (i.e. span size) will be balanced and varied within and across participants through an adaptive staircase procedure. Two different difficulty thresholds (80% and 99%) will be determined in blocks at the beginning of the experiment. These difficulty thresholds will be used to investigate how task difficulty affects performance. Altogether, this study will attempt to reconcile debates over the differential contributions of working memory in mental arithmetic.

Methods

Participants

All research will be performed in accordance with the ethical standards of the Institutional Review Board. Written informed consent will be obtained from all participants and

they will either be given course credit through the Human Subjects Lab Pool or be reimbursed $30 for their participation. We used the software program G*Power to conduct a power analysis (Faul et al., 2009). F statistics or η_p^2 values for the interaction between WM load and arithmetic operation could not be derived from Lee & Kang (2002) nor Cavdaroglu & Knops (2016). However, other 2- and 3- way interactions were provided from Imbo & LeFevre (2010) (e.g. culture × problem difficulty; culture × problem difficulty × presentation format) to approximate values for the power analysis. Our goal was to obtain .90 power to detect a partial eta-squared (η_p^2) of .07 for a 3-way interaction at alpha = .05. We used the η_p^2 reported for the 3-way interaction between culture x problem difficulty x presentation format in Imbo & LeFevre experiment 2 (2010), as this was the most conservative effect size reported relating to arithmetic performance. For the statistical test, we chose "ANOVA: Repeated measures, within-between interaction" because the interaction from Imbo & LeFevre (2010) contained within factors (problem difficulty & and presentation format) and a between factor (culture). We input the reported $\eta_p^2 = .07$ after clicking "Determine =>". Calculating this provided an effect size of 0.27. The assumed correlation between repeated measures was left at the default of 0.5 because we had no other underlying assumptions about the repeated measures. In addition, we specified that there were 2 groups (Chinese and U.S. math educated students) and 16 measurements (i.e. 2 arithmetic operation × 2 difficulty × 4 WM load types). While four factors are present here, our main focus is the 2-way interaction between operation and WM load. The additional factors used in the G*Power analysis helped derive a more conservative estimate for the number of participants needed and will be used in subgroup analyses explained further below. Following these specifications, a minimum of 14 participants is required to be powered to detect an interaction similar to that in Imbo & LeFevre (2010). Prior meta-analytic data also suggests the

average sample size among dual-task arithmetic experiments (containing both within and between designs) has around 20 participants with a range from 10-60. Following prior literature and our power analyses, we plan on collecting data from a sample larger than any other study before. As such, 100 participants would be sufficiently powered to detect our key interaction within our main model and secondary analyses. The final sample will include both English and native Chinese speaking participants from the UC Irvine Human Subjects Lab Pool and from online advertisements. Ideally, we want an equal proportion of Chinese and non-Chinese educated participants, but this is not guaranteed in recruitment. We will not analyze data until data collection has been completed. Participants may drop the experiment at any time between sessions 1 and 2 for whatever reason but are incentivized to complete both sessions to receive course credit or monetary compensation. All participants will have normal or corrected-to-normal vision.

Stimuli

All tasks used in these experiments were created through PsychoPy 3 (Peirce et al., 2019). Performance on both span tasks and arithmetic operations will be measured in reaction time (RTs in ms) and accuracy (ACCs in percentage correct). For examples of the tasks, see Figures 1 & 2. Arithmetic problems used in this experiment are the same as in Cavdaroglu & Knops (2016). Working memory staircase tasks are based on the descriptions used in Cavadaroglu & Knops (2016). All materials including experimental tasks and protocol used for the registered report will be available online as supplementary materials.

Subtraction

Subtraction problems will be presented using a 2-alternative forced choice (2AFC) paradigm. Participants will be presented with simple two-digit – two-digit problems for 2 s.

There will be no borrowing or crossing of decade boundaries to minimize central executive involvement. Participants will then choose from two answer choices which will be displayed for 3 s or until participants respond. Three different sets of subtraction problems will be used across three rounds (round 1: subtraction only; round 2: subtraction under phonological load; round 3: subtraction under visuospatial load) with easy and hard working memory loads split across 2 blocks. The order of the three sets as well as the difficulty blocks will be counterbalanced across participants. Each set will contain 28 different subtraction problems where each will be displayed twice in total with a different answer pair each time. The order of the three sets will be counterbalanced across all participants. In half of the answer pairs, the correct and alternative answers will have a distance of 2; whereas the other half will have a distance of 10. Distance from correct response can be either in the positive or negative direction. For example, for the problem 36-14, the two answer pairs are 12 vs. 10 (distance = -2) or 12 vs. 22 (distance = +10). Problems with a decade in one of the operands or in the result were excluded. Eleven was not used as an operand.

Multiplication

Multiplication problems will be presented using a 2AFC paradigm. Participants will be presented with simple one-digit by one-digit and two-digit by one-digit multiplication problems. Participants will then choose from two answer alternatives which will be displayed for 3 s or until participants respond. Three different sets of multiplication problems will be used across three rounds of tasks (round 1: multiplication only; round 2: multiplication under phonological load; round 3: multiplication under visuospatial load) with easy and hard working memory loads split across 2 blocks. The order of the three sets as well as the difficulty blocks will be counterbalanced across participants. Each set will contain 28 different subtraction problems

where each will be displayed four times in total with a different answer pair each time. Among the four answer pairs, one contained a response alternative from the multiplication table of the first operand, another contained an alternative from the multiplication table of the second operand (table-related response alternatives) and the other two pairs contained response alternatives that are not from either operand's multiplication table (non-table-related response alternatives). For example, for the problem 12 × 7, the four different answer pairs were 84 vs 98 (98 from 7's table), 84 vs 72 (72 from 12's table), 84 vs 64, and 84 vs 94. Half of the problems will be two-digit by one-digit and the other half will be one-digit multiplication. In one-digit multiplication trials, the smaller operand will precede the larger operand. In two-digit by one-digit trials, the two-digit operand will precede the one-digit. The two-digit number will be smaller than twenty. The one-digit number will be larger than two. Tie problems (e.g. 6 × 6) and problems with a decade in the operand or result will be excluded. Products are all below 100 to restrict responses to be two-digits at most like in the subtraction task.

Phonological staircase

Following the same task designs as those outlined in Cavdaroglu & Knops (2016), participants' phonological processing span will be measured using an adaptive staircase procedure of letter sequences. Participants will be instructed to keep a sequence of letters – in original order – in mind and decide whether a second set of letters (shown 7s after onset of the first sequence) contained the exact same order of letters or not. Letter sequences will be displayed for a duration of 0.4 s * n – n indicating number of letters – followed by 3 s on a fixation screen before participants are given 4 s to respond. Participants will be presented upper case letters in the first sequence and tested using lowercase letters (B C D vs. b c d) in order to encourage participants to use their phonological rather than visual memory. In half of the trials,

the test sequence will have the same letters in the exact order as the first sequence (e.g. ‘b c d’ and ‘b c d’); whereas in the other half of the trials the position of two letters will be swapped (e.g. ‘b c d’ and ‘b d c’). The ‘F’ and ‘J’ keys will be used for responding to allow for natural hand placement on the keyboard. The task will start with 3 letters and reach a maximum of 9 letters and a minimum of 1 letter. After three correct responses in a row, the difficulty of the task will be increased by 1 letter otherwise, if there are three consecutive incorrect responses, the difficulty of task will be decreased by 1 letter until the minimum number of letters are reached or until a correct response is given. 30 trials will be conducted to measure phonological span. In addition, a Weibull function will be fit on the data where the inverse of the Weibull function will be used to determine the number of letters corresponding to 80 and 99% accuracy. The two threshold levels were chosen to examine the effect of task difficulty (low vs high) on arithmetic performance in both single- and dual-task conditions. In each trial, the string of letters will be randomly chosen from this set of 10 consonants [B, C, D, F, G, H, J, K, L, M]. Vowels were excluded to prevent use of semantic strategies and other consonants were excluded to maintain the same number of digits to letters. In total, the staircase will contain 30 trials.

Visuospatial staircase

The visuospatial span task will also follow similar outlines to those in Cavdaroglu & Knops (2016), where span will be measured using an adaptive staircase procedure on dot-matrices. Participants will be instructed to keep the locations of dots within a 5×5 grid in mind and decide if a second grid (shown 7s after onset of the first grid) contained the exact same locations of dots. Dot-arrays will be displayed for a duration of 0.4 s * n – n indicating number of dots – followed by 3 s on a fixation screen before participants are given 4 s to respond. In half of the trials, the position of the test dots will be in the same position; whereas in the other half of

the trials, the position of two dots will be replaced somewhere else on the grid. The 'F' and 'J' keys will be used for responding. The task will start with 3 dots and reach a maximum of 9 dots and a minimum of 1 dot After three correct responses in a row, the difficulty of the task will be increased by 1 dot otherwise, if there are three consecutive incorrect responses, the difficulty of the task will be decreased by 1 dot until the minimum number of dots are reached or until a correct response is given. 30 trials will be conducted to measure visuospatial span. Finally, a Weibull function will be used to determine the 80 and 99% accuracy thresholds for the dual-task condition.

Procedure

The study will be a 2×3 factorial design using within-subject factors. The within-subject factors will be arithmetic operation type (subtraction or multiplication) and WM load type (no load, PL load, and VSSP load). No load (i.e. arithmetic alone) conditions will serve as controls against dual-task conditions. While culture and difficulty will be part of the analysis, these will only be considered in the subgroup analyses and not for additional interactions, because our focus is on the operation × load interaction. The entire experiment will be conducted online through video conferencing in which an experimenter will guide the participant in downloading the required materials and protocol for completing experimental tasks. The experiment will be administered within two sessions that will be scheduled to be around the same time and spread apart by 1 week. Participants will also be instructed to abstain from taking any alcohol or drugs prior to either session. Participants will complete the experiment using their own devices. To ensure that reactions times are accurate and mostly consistent across different devices and operating systems, participants will be instructed to use either a home desktop or laptop rather than a tablet or mobile phone. No information related to the participants' devices, such as IP

address, will be maintained except for the operating system (e.g. Windows 10, Mac-OS) in order to ensure proper installation of PsychoPy and the experiment itself. Recordings will also not be taken to respect the privacy of the participants.

In session 1, participants will be given a brief questionnaire regarding demographic information and math education background before being introduced to the PsychoPy environment and to downloading the experimental tasks. These questions will include asking about their current major and the number of math courses they have taken since entering university. In addition, we will ask specific math background questions including, "Prior to coming to university, in which country did you receive the majority of your math education?", "If you were taught how to use an abacus or mental abacus strategy for doing math, how often have you used it? (Never taught; Never used; Rarely; Sometimes; Often; Very often)", and "Do you consider yourself an A, B, C, D, or F student compared to your peers?". Altogether, these questions will allow us to potentially examine differences in math proficiency among our sample, especially in our comparison between the Chinese-educated student group and the non-Chinese-educated student group. From here, participants will be given the adaptive phonological and visuospatial staircase tasks. Prior to the staircase, 10 practice trials will be administered to familiarize the participant with the stimuli and testing environment. Discounting the practice trials, there will be 30 trials per staircase for a total of 60 trials to determine difficulty thresholds. The order of these tasks will be randomized and counterbalanced for all participants. Staircase performance from session 1 will be used to determine easy and hard span levels for the dual-task conditions used in session 2. In total, the first session will take approximately 60 minutes.

In session 2, participants will start the dual-task experiment. Participants will download their PsychoPy tasks that have been modified to fit their difficulty levels. Participants will then

complete arithmetic alone and under load over 4 experimental blocks (multiplication-easy load, multiplication-hard load, subtraction-easy load, subtraction-hard load). The order of these tasks will follow a block-randomization wherein the single-arithmetic task will always go first in the block followed by either the visuospatial or phonological loads. Half of the participants will receive the visuospatial load before the phonological load while the other half will receive the phonological load first. The order of the four blocks has also been randomized and counterbalanced for each participant such that each of the possible sequences as well as their reverse orders appear an equal number of times. 10 practice trials will be given before the start of the first block to familiarize participants with the dual-task procedure. Participants will then complete each block which will contain 28 arithmetic problems for each condition (arithmetic alone, with PL load, with VSSP load) for a total of 336 trials. The order of conditions will also be randomized and counterbalanced. At the end of each block, participants will be given up to a 5-minute break. Participants will finish after completing the 4^{th} block. In total, the second session should take no more than 2 hours to complete.

Analysis plan

In this experiment, we will focus on the key interaction predicted by Lee & Kang (2002). Specifically, we will test the following hypotheses:

Hypothesis 1: As predicted by Lee & Kang (2002), there is an interaction between arithmetic operation type and WM load type; specifically:

Hypothesis 1a: Multiplication performance is slower and less accurate under PL load compared to VSSP load

Hypothesis 1b: Subtraction performance is slower and less accurate under VSSP load compared to PL load.

In addition to these, we will test secondary hypotheses regarding the differences between single-task arithmetic conditions vs each of the dual-task conditions as they were reported in Lee & Kang (2002) such that:

Hypothesis 1c: Multiplication performance alone is significantly faster than under PL load but not VSSP load.

Hypothesis 1d: Subtraction performance alone is significantly faster than under VSSP load but not PL load.

According to Imbo & LeFevre (2010), the crossover effect may be found within Chinese-educated samples; but not US-educated samples, thus we will test the following hypotheses:

Hypothesis 2: Receiving primary math education from China but not the US will lead to differences in load type by arithmetic operation performance, specifically:

Hypothesis 2a: Multiplication performance is slower and less accurate under PL load compared to VSSP load only in Chinese samples.

Hypothesis 2b: Subtraction performance is slower and less accurate under VSSP load compared to PL load only in Chinese samples.

Hypothesis 2c: Multiplication performance alone is significantly faster than under PL load but not VSSP load only in Chinese samples.

Hypothesis 2d: Subtraction performance alone is significantly faster than under VSSP load but not PL load only in Chinese samples.

In order to test these hypotheses 1a-1d, we will conduct multiple 2 × 2 ANOVAs under four model specifications (for summary of planned analyses, see Table 2). The first model will include all participants and both difficulty levels. We will then test the robustness of this interaction effect by restricting the data in the following three ANOVA models. The second

model will restrict the sample to only those students who received their primary math education in China. We will obtain this information using the questionnaire mentioned in the procedures of session 1. This second model is predicated on hypotheses 2a-2d in which we will first run a 2 × 2 × 2 ANOVA with the Chinese vs non-Chinese samples as the between subject factor and load and operation as the within subject factors. While we are investigating this possible group difference, the crossover interaction is our main interest. Unequal sample sizes may be an issue for the Chinese vs non-Chinese model, so we will run a Tukey-Kramer test as a post-hoc adjustment. The third and fourth models will then restrict the data to include only easy WM load tasks or only hard WM load tasks, respectively. If any of the above models produce a significant interaction effect, we will conduct post hoc analyses to see if results align with hypotheses 1a - 2d. One additional secondary analysis will investigate possible differential sequence effects. This model will test whether the crossover interaction is observed for the first presented arithmetic operation under load (Table 2: last column), for which performance should be less prone to order effects.

Testing these multiple hypotheses would invariably inflate the probability of type-1 errors. However, we choose not to adjust error levels for each statistical test, because a statistically significant interaction does not guarantee any of the more specific hypotheses to be supported. How closely our results align with our predictions will show different levels of evidence for this theorized crossover effect and predicted simple effects. For hypotheses 1a-1d, we will conclude that there is strong support for the underlying theory if we detect an interaction and main load effects in directions consistent with Lee & Kang (2002) within our main specifications containing all participants. We will conclude there is mixed evidence for the crossover effect if only one of the main load effects are consistent with predictions within the

main model (i.e., a) if VSSP affects subtraction but not multiplication or b) PL affects multiplication but not subtraction, but not both a and b) or if we can only find the interaction only in one or more of the subgroup analyses; for example, if the crossover effect is only present in the Chinese sample but not the US sample or only in hard but not easy load conditions. If results are fully null, we will conclude that we were unable to find evidence for an interaction. Results of analyses will be reported regardless of whether our hypotheses are supported or not.

As a complement to the frequentist analyses of the interaction effect, we will also report a Bayesian analysis of this effect for the main model (whole group) to examine the relative support for both our hypotheses of interest and the null hypothesis. We will conduct a Bayesian repeated measures ANOVA, dependent on the 2 × 2 factors in the main model. Following Morey & Rouder (2011), we will set a non-informative Jeffreys prior width of 0.5 to correspond to a small effect. Such analyses result in a Bayes factor (BF_{10}), which can be interpreted as the likelihood ratio for the alternative hypothesis over the null. Given that the Bayes factor (BF_{10}) is a ratio of the likelihood for the alternative hypothesis over the null hypothesis, the inverse of the Bayes factor (BF_{01}) can be interpreted as the likelihood ratio for evidence of the null hypothesis over the alternative hypothesis. Following Jeffreys (1961) we will use the following designations to interpret the strength of the Bayes factors: 0–3 offer anecdotal support for H1, 3–10 moderate support for the H1, 10–30 strong support for H1, 30–100 very strong evidence for H1, and values greater than 100 offer decisive evidence for H1. Conversely, we use the inverse of these ranges to interpret support for the null hypothesis (BF_{01} anecdotal 0.33–0, moderate 0.10–0.33, strong 0.10–0.03, very strong 0.03–0.01).

Data will be analyzed using RStudio (RStudio Team, 2020), specifically tidyverse for data visualization (Wickham et al., 2019), rstatix for ANOVAs (Kassambara, 2021), ggplot2

(Wickham, 2016), and ggpubr (Wickham, 2020) for publication-ready figures and tables. The RMarkdown file will be available as supplementary material to reproduce analyses. Where appropriate, Holm-Bonferroni correction will be used to correct for multiple comparisons in post-hoc testing (Holm, 1979). Huynh–Feldt correction will be used if sphericity is violated. Bayesian analyses will be conducted using the repeated measures ANOVA function in JASP (JASP Team, 2020). All reaction time (RT) analyses will be based on correct trials only. Accuracy or response times outside the range of a participant's mean ± 3 SDs will be discarded from further analyses. Responses faster than 200 ms will also be discarded. All data will be made publicly available after data analysis has finished.

Appendix

Table 1. Studies that tested arithmetic operation × WM load type interaction

Author	sample size	WM tasks	Arithmetic tasks	PL load effect (multiplication vs subtraction) *d*, CI	VSSP load effect (subtraction vs multiplication) *d*, CI
Lee, K. M., & Kang, S. Y. (2002)	10	Repeat nonword string (PL), Matching abstract shapes and location (VSSP)	exact subtraction, exact multiplication	2.39 [1.24, 3.54]	3.35 [1.99, 4.71]
Imbo, I., & LeFevre, J.A. (2010) – Canadian sample	57	Repeat nonword string (PL), 4×4 grid location task (VSSP)	two-digit subtraction, one × two-digit multiplication	0.11 [-.40, .63]	-0.13 [-.65, .39]
Imbo, I., & LeFevre, J.A. (2010) – Chinese sample	73	Repeat nonword string (PL), 4×4 grid location task (VSSP)	two-digit subtraction, one × two-digit multiplication	-0.05 [-.50, .41]	-0.04 [-.50, .42]
Cavdaroglu, S., & Knops, A. (2016)	32	Letter span (PL), 5×5 grid location task (VSSP)	2AFC multiplication (one × one; two × one digit), 2 AFC subtraction (two - one digit)	0.10 [-.59, .8]	0.00 [-.70, .69]
Chen, E.H., Jaeggi, S.M., & Bailey, D.H. – Chinese sample	100	Letter span (PL), 5×5 grid location task (VSSP)	2AFC multiplication (one × one; two × one digit), 2 AFC subtraction (two - one digit)		
Chen, E.H., Jaeggi, S.M., & Bailey, D.H. – U.S. sample	100	Letter span (PL), 5×5 grid location task (VSSP)	2AFC multiplication (one × one; two × one digit), 2 AFC subtraction (two - one digit)		

Note. Cohen's *d* were calculated for the last two columns. *d* represents effect size between multiplication and subtraction RT performance under respective (PL or VSSP) load.

Table 2

Planned ANOVA on arithmetic accuracy and reaction time by model specification

	Whole sample			Chinese only			Easy only			Hard only			First Under Load		
Factor	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2	F	p	η_p^2
PL vs VSSP × Multiplication															
PL vs VSSP × Subtraction															

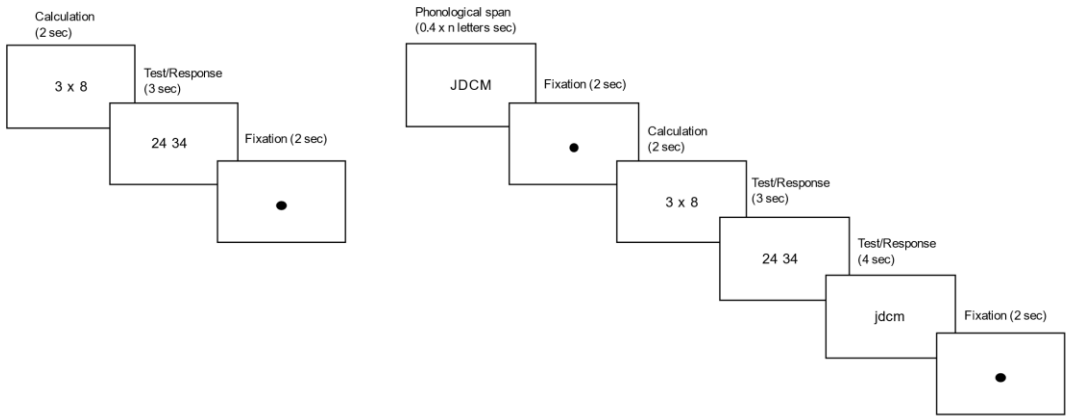

Figure 1. Single multiplication task (left). Dual-task multiplication with phonological letter WM load (right)

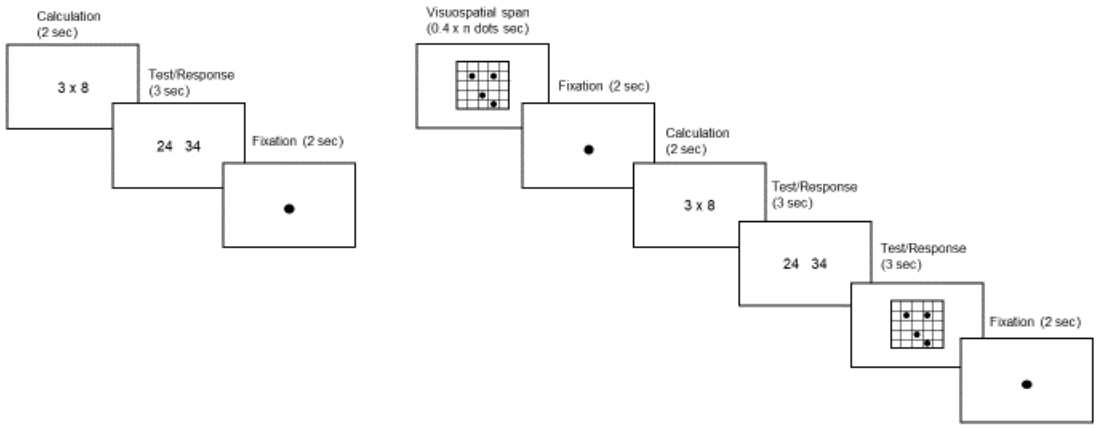

Figure 2. Single multiplication task (left). Dual-task multiplication with visuospatial WM load (right)

References

Cavdaroglu, S., & Knops, A. (2016). Mental subtraction and multiplication recruit both phonological and visuospatial resources: Evidence from a symmetric dual-task design. *Psychological Research*, *80*(4), 608-624.

Chen, E. H., & Bailey, D. H. (2021). Dual-task studies of working memory and arithmetic performance: A meta-analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(2), 220.Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex*, 33(2), 219-250.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, *20*(3-6), 487-506.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149-1160.

Imbo, I., & LeFevre, J. A. (2010). The role of phonological and visual working memory in complex arithmetic for Chinese-and Canadian-educated adults. *Memory & Cognition*, *38*(2), 176-185.

Kawashima, R., Taira, M., Okita, K., Inoue, K., Tajima, N., Yoshida, H., ... & Fukuda, H. (2004). A functional MRI study of simple arithmetic—a comparison between children and adults. *Cognitive Brain Research*, *18*(3), 227-233.

Lee, K. M. (2000). Cortical areas differentially involved in multiplication and subtraction: a functional magnetic resonance imaging study and correlation with a case of selective

acalculia. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, *48*(4), 657-661.

Lee, K. M., & Kang, S. Y. (2002). Arithmetic operation and working memory: Differential suppression in dual tasks. *Cognition*, *83*(3), B63-B68.

Navon, D., & Miller, J. (1987). Role of outcome conflict in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 435.

Pashler, H. (1994). Graded capacity-sharing in dual-task interference?. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(2), 330.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behavior Research Methods*. 10.3758/s13428-018-01193-y

Prado, J., Mutreja, R., Zhang, H., Mehta, R., Desroches, A. S., Minas, J. E., & Booth, J. R. (2011). Distinct representations of subtraction and multiplication in the neural systems for numerosity and language. *Human Brain Mapping*, *32*(11), 1932-1947.