Journal of Numerical Cognition

PsychOpen GOLD

Empirical Research [Registered Report]

Check for updates

No Clear Support for Differential Influences of Visuospatial and Phonological Resources on Mental Arithmetic: A Registered Report

Edward H. Chen [1], Drew H. Bailey [1], Susanne M. Jaeggi [1]

[1] *School of Education, University of California, Irvine, CA, USA.*

Journal of Numerical Cognition, 2022, Vol. 8(2), 259–280, https://doi.org/10.5964/jnc.8421

Received: 2022-02-21 • **Accepted:** 2022-05-24 • **Published (VoR):** 2022-07-28

Handling Editor: André Knops, Université de Paris, Paris, France

Corresponding Author: Edward H. Chen, School of Education, University of California, 3200 Education Bldg, Irvine, CA 92697, USA. E-mail: cheneh1@uci.edu

Supplementary Materials: Data, Materials, Preregistration [see Index of Supplementary Materials]

Abstract

Several working memory processes have been 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. A previous dual task experiment produced these effects, but subsequent experiments have yielded inconsistent results. Because the reasons for these inconsistencies are not immediately apparent, the current study systematically reviewed these subsequent attempts and attempted to replicate this effect in a within-subjects dual task experiment using tasks developed from prior work across a number of different subsamples. In contrast to the original finding, we observed no differential impact of specific working memory secondary tasks by arithmetic operation in any of our analyses. However, our analyses do not entirely rule out the possibility of differential effects of working memory tasks. Our findings suggest that the working memory facet by arithmetic operation interactions observed in previous work may be idiosyncratic in nature and difficult to predict a priori in subsequent experiments.

Keywords

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 and 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 A1). 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 and Kang (2002) study. Strikingly, while there is variation in the kinds of tasks and samples among others, the original study does not seem to be sufficiently different in its design that would lead to the discrepancy in effect sizes (Table A1). Neither of the two partial replications used an entirely within-subject design like Lee and Kang (2002), which could have potentially led to the discrepancy in effect sizes. The current paper will go beyond Lee and Kang (2002) and 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 and LeFevre (2010) attempted to replicate the findings using a mix of native Chinese and Canadian participants to perform arithmetic problems under load (see Table A1 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 and 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 and 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 and Knops (2016) attempted to resolve this issue by having German participants perform arithmetic under similar load conditions to Lee and 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,

1) Cohen's d was calculated by hand. RT means were taken from reported values within Lee and Kang (2002) 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: $d = (M_1 - M_2)/\sigma$, where σ is the pooled standard deviation of the two means: pooled $SD = \sqrt{(SD_1^2 + SD_2^2)/2}$. The same method was used to calculate the effect size for visuospatial load.

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, 2000; Lee & Kang, 2002), 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 and 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 and 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 and 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). Given that there are several ways to probe potential interactions in dual task arithmetic experiments, the robustness of these findings warrants further testing. 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 and 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 and Knops (2016) improved upon the original design of Lee and Kang (2002), some remaining issues need to be experimentally investigated. The current design will go beyond Cavdaroglu and 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 tested 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 and 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 was 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 and Kang (2002) can be replicated using similar procedures and tasks as used by Cavdaroglu and Knops (2016). Participants solved 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 and Kang (2002). However, attempts to replicate this large dissociation since have not been wholly successful (see Table A1). Task difficulty (i.e., span size) was balanced and varied within and across participants through an adaptive staircase procedure. Two different difficulty thresholds (80% and 99%) were determined in blocks at the beginning of the experiment in session 1. These difficulty thresholds were 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 and provide insight with respect to potential underlying mechanisms related to mathematical cognition.

Method

Participants

Power Analysis

We used the software program G*Power to conduct an a priori 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 and Kang (2002) nor Cavdaroglu and Knops (2016). However, other 2- and 3-way interactions were provided from Imbo and 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 × problem difficulty × presentation format in Imbo and 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 and LeFevre (2010) contained within factors (problem difficulty & and presentation format) and a between factor (culture). We inputted 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 in our design, our main focus was 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 was required to be powered to detect an interaction similar to that in Imbo and LeFevre (2010) and our design had an estimated power of 0.94. Prior meta-analytic data also suggests the average sample size among dual-task arithmetic experiments (containing both within and between designs) consists of around 20 participants with a range from 10-60. Following prior literature and our power analyses, we planned on collecting data from a sample larger than any other study before. As such, we determined that 100 participants would be sufficiently powered to detect our key interaction within our main model and secondary analyses.

Following this plan, we recruited and ran 100 total participants from the University of California, Irvine (Female = 64, age range = 18 – 25 years old, $M = 20.1$ ($SD = 1.3$). 22 of the final analysis sample received the majority of math education in China prior to entering university studies in the US. All participants had normal or corrected-to-normal vision. All research was performed in accordance with the ethical standards of the Institutional Review Board. Written informed consent was obtained from all participants and were given course credit through the Human Subjects Lab Pool or were reimbursed $30 for their participation.

Stimuli

All tasks used in these experiments were created through PsychoPy 3 (Peirce et al., 2019). Performance on the span tasks and arithmetic will be measured by reaction time (RTs in ms) and accuracy (ACCs in percentage correct). For examples, see Figure A1. Arithmetic problems used in this experiment are the same as in Cavdaroglu and Knops (2016). Working memory staircase tasks are based on the descriptions used in Cavdaroglu and Knops (2016). Strategy report is a one-item survey question asking about strategy use. All materials including experimental tasks and protocol used are available online as Supplementary Materials.

Subtraction

Subtraction problems were presented using a 2-alternative forced choice (2AFC) paradigm. Participants were presented with simple two-digit – two-digit problems for 2 s. There were no borrowing or crossing of decade boundaries to

minimize central executive involvement. Participants then chose from two answer choices which were displayed for 3 s or until participants respond. Three different sets of subtraction problems were 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 were counterbalanced across participants. Each set contained 28 different subtraction problems where each was displayed twice in total with a different answer pair each time. The order of the three sets was counterbalanced across all participants. In half of the answer pairs, the correct and alternative answers had a distance of 2; whereas the other half had a distance of 10. This was done in order to encourage participants to take into account both decades and units and to discourage the strategy of paying attention only to the units or decades. Distance from correct response were either in the positive or negative direction. For example, for the problem 36-14, the two answer pairs were 22 vs. 20 (difference = -2) or 12 vs. 22 (difference = +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 were presented using a 2AFC paradigm. Participants were presented with simple one-digit by one-digit and two-digit by one-digit multiplication problems. Participants then chose from two answer alternatives which were displayed for 3 s or until participants responded. Three different sets of multiplication problems were 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 was counterbalanced across participants. Each set contained 28 different subtraction problems where each was 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 were 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 were two-digit by one-digit and the other half were one-digit multiplication. In one-digit multiplication trials, the smaller operand preceded the larger operand. In two-digit by one-digit trials, the two-digit operand preceded the one-digit. The two-digit number was smaller than twenty. The one-digit number was larger than two. Tie problems (e.g., 6 × 6) and problems with a decade in the operand or result were excluded. Products were 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 and Knops (2016), participants' phonological processing span was measured using an adaptive staircase procedure of letter sequences. Participants were instructed to keep a sequence of letters – in original order – in mind and decide whether a second set of letters (shown 7 s after onset of the first sequence) contained the exact same order of letters or not. Letter sequences were 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 were 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 had 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 were swapped (e.g., 'B C D' and 'b d c'). The 'F' and 'J' keys were 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 increased by 1 letter otherwise, if there were three consecutive incorrect responses, the difficulty of task decreased by 1 letter until the minimum number of letters are reached or until a correct response is given. 30 trials were conducted to measure phonological span. In addition, a Weibull function was fit on the data where the inverse of the Weibull function was 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 was 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 contained 30 trials.

Visuospatial Staircase

The visuospatial span task also followed similar procedures to those used in Cavdaroglu and Knops (2016), where span was measured using an adaptive staircase procedure on dot-matrices. Participants were instructed to keep the locations of dots within a 5×5 grid in mind and decide if a second grid (shown 7 s after onset of the first grid) contained the exact same locations of dots. Dot-arrays were 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 positions of the test dots were in the same position; whereas in the other half of the trials, the positions of two dots were replaced somewhere else on the grid. The 'F' and 'J' keys were used for responding. The task started with 3 dots and reached a maximum of 9 dots and a minimum of 1 dot. After three correct responses in a row, the difficulty of the task increased by 1 dot otherwise, if there were three consecutive incorrect responses, the difficulty of the task decreased by 1 dot until the minimum number of dots was reached or until a correct response was given. 30 trials were conducted to measure visuospatial span. Finally, a Weibull function was used to determine the 80 and 99% accuracy thresholds for the dual-task condition.

Procedure

The study used a 2×3 factorial design using within-subject factors. The within-subject factors were arithmetic operation type (subtraction or multiplication) and WM load type (no load, PL load, and VSSP load). No-load (i.e., arithmetic alone) conditions served as controls against dual-task conditions. While culture and difficulty were part of the analysis, these were only considered in the subgroup analyses and not for additional interactions, because our main focus was on the operation × load interaction. The entire experiment was conducted online through video conferencing in which an experimenter guided the participant in downloading the required materials and protocol for completing experimental tasks. The experiment was administered within two sessions that were scheduled to be around the same time and spread apart by one week. Participants were also instructed to abstain from taking any alcohol or drugs prior to either session. Participants completed the experiment using their own devices. To ensure that reactions times were sufficiently accurate and consistent across different devices and operating systems, participants were 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, were 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 were also not taken to respect the privacy of the participants.

In session 1, participants were given a brief questionnaire to capture their demographic information and math education background before being introduced to the PsychoPy environment and to downloading the experimental tasks. These questions included asking about their current major and the number of math courses they have taken since entering university. In addition, we asked 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 allowed 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 were given the adaptive phonological and visuospatial staircase tasks. Prior to the staircase, 10 practice trials were administered to familiarize the participant with the stimuli and testing environment. Discounting the practice trials, there were 30 trials per staircase for a total of 60 trials to determine difficulty thresholds. The order of these tasks were randomized and counterbalanced for all participants. Staircase performance from session 1 were used to determine easy and hard span levels for the dual-task conditions used in session 2. In total, the first session took approximately 60 minutes.

In session 2, participants started the dual-task experiment. Participants downloaded their PsychoPy tasks that were modified to fit the appropriate difficulty levels as determined in session 1. Participants then completed 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 followed a block-randomization wherein the single-arithmetic task was always administered first in the block followed by either the visuospatial or phonological loads. Half of the participants received the visuospatial load before the phonological load, while the other half received the phonological load first. The order of the four blocks was also randomized and counterbalanced for each participant such that each of the possible sequences as well as their reverse orders appeared an equal number of times. 10 practice trials were given before the start of the first block to familiarize participants with the dual-task procedure. Participants then completed each block which contained 28 arithmetic problems for each condition (arithmetic alone, with PL load, with VSSP load) for a total of 336 trials. The order of conditions was also randomized and counterbalanced. At the end of each block, participants were be given up to a 5-minute break. Participants finished after completing the 4^{th} block. In total, the second session took no more than 2 hours to complete.

Analysis Plan

In this experiment, we focused on the key interaction predicted by Lee and Kang (2002). Specifically, we tested the following hypotheses:

> Hypothesis 1: As predicted by Lee and Kang (2002), we expected 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 tested secondary hypotheses regarding the differences between single-task arithmetic conditions vs. each of the dual-task conditions as they were reported in Lee and 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 and LeFevre (2010), the crossover effect may be found within Chinese-educated samples; but not US-educated samples, thus we tested the following hypotheses:

> Hypothesis 2: Receiving primary math education from China but not the US is associated with 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-educated samples.
>
> Hypothesis 2b: Subtraction performance is slower and less accurate under VSSP load compared to PL load only in Chinese-educated samples.
>
> Hypothesis 2c: Multiplication performance alone is significantly faster than under PL load but not VSSP load only in Chinese-educated samples.
>
> Hypothesis 2d: Subtraction performance alone is significantly faster than under VSSP load but not PL load only in Chinese-educated samples.

In order to test Hypotheses 1a-1d, we conducted multiple 2×2 ANOVAs under four model specifications (for summary of planned analyses, see Table A2). The first model included all participants and both difficulty levels. We then tested the

robustness of this interaction effect by restricting the data in the following three ANOVA models: easy load condition trials only, hard load condition trials only, and first block trials only. The first block model tested whether the crossover interaction was observed for the first presented arithmetic operation under load (Table A2: last column), for which performance was assumed to be less prone to order effects. To test Hypotheses 2a-2d, we restricted the sample to only those students who reported having received the majority of their math education in China prior to entering university. We conducted both a 2×2 ANOVA of the restricted sample and compared the Chinese-educated students to the rest of our sample using a 2×2×2 ANOVA with country of primary math education as a between subject factor. While we investigated this possible group difference, the crossover interaction was our main interest. Given the unequal sample sizes in the Chinese vs. non-Chinese model, we ran a Tukey-Kramer test as a post-hoc adjustment. If any of the above models produced a significant interaction effect, we conducted post hoc analyses to test whether results aligned with Hypotheses 1a-2d.

Even though we acknowledge that testing these multiple hypotheses inflates the probability of type-1 errors, we chose 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. Instead, we reported on the level of support for the theorized crossover effect and predicted simple effects based on how closely our reported findings aligned with our predictions. For Hypotheses 1a-1d, we concluded that there was strong support for the underlying theory if we detected an interaction and main load effects in directions consistent with Lee and Kang (2002) within our main specifications containing all participants. We concluded there was mixed evidence for the crossover effect if only one of the main load effects was 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 only found the interaction in one or more of the subgroup analyses; for example, if the crossover effect was only present in the Chinese-educated sample but not the US sample or only in hard but not easy load conditions. If results were fully null, we concluded that we were unable to find evidence for an interaction. Results of analyses will be reported regardless of whether our hypotheses were supported or not.

As a complement to the frequentist analyses of the interaction effect, we also report a Bayesian analysis for the main model (whole group) to examine the relative support for both our hypotheses of interest and the null hypothesis. We conducted a Bayesian repeated measures ANOVA, dependent on the 2×2 factors in the main model. Following Morey and Rouder (2011), we 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 used 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 were analyzed primarily in JASP using its frequentist and Bayesian repeated measures ANOVA and paired-sample *t*-test functions (JASP Team, 2020). Data were organized for JASP using RStudio (RStudio Team, 2020), specifically tidyverse for data visualization and formatting (Wickham et al., 2019). The RMarkdown is available in the Supplementary Materials to reproduce data created for JASP. Where appropriate, Holm-Bonferroni correction was used to correct for multiple comparisons in post-hoc testing (Holm, 1979). Huynh–Feldt correction was used when sphericity was violated. Bayesian analyses were conducted using the Bayesian repeated measures ANOVA function in JASP (JASP Team, 2020). All reaction time (RT) analyses were based on correct trials only. Accuracy or response times outside the range of a participant's mean ± 3 *SD*s were discarded from further analyses. Responses faster than 200 ms were also discarded. Based on that criterion, 1.02% of trials in single arithmetic blocks and 3.56% of the trials in dual-task blocks were eliminated. In addition, 3 participants were excluded from data analyses for not responding in a majority of trials during the second session. All data are publicly available in PsychArchives (see Supplementary Materials). Of note, even though our participants were tested at home on their own devices, average reaction times per WM load condition within our study were comparable to those found in Lee and Kang (2002) and Cavdaroglu and Knops (2016) (Table A1).

Deviations in Pre-Registration Analyses

The following analyses were either changed or added from the pre-registration. Full documentation of all deviations can be found in a document within the Supplementary Materials. The 2×2×2 ANOVA investigating the differential effect of WM load on arithmetic operation between the samples receiving education from the US and China was included in the pre-registration, but we also included the 2×2 ANOVA analyses which only looked at the Chinese-educated subsample as an additional robustness test. We also conducted additional Bayesian paired samples *t*-tests in addition to the Bayesian repeated measures ANOVA to investigate post-hoc differences in reaction time and accuracy for Hypotheses 1a, 1b, 2a, and 2b. Conclusions did not vary across methods.

Results

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

In contrast to our hypothesis, in the full sample, multiplication performance was not significantly slower (Figure A2) nor was it less accurate (Figure A3) under PL load compared to VSSP load. ANOVA results from Tables A4 and A5 yielded no significant difference in multiplication reaction time [RT: $F(1, 96) = 1.20$, $p = .28$, $\eta_p^2 = .01$] nor accuracy [ACC: $F(1, 96) = 0.49$, $p = .49$, $\eta_p^2 = .01$] between verbal and visuospatial dual-task load. We ran complementary Bayesian *t*-tests of WM load on multiplication RT and ACC for the full sample. We found stronger evidence for the null hypothesis than for Hypothesis 1a such that there was no difference in multiplication RT and ACC by WM load type (RT: BF_{01}: 4.99; ACC: BF_{01}: 7.03). Higher BF_{01} indicate greater support for the null hypothesis over the alternative. In addition to our Bayesian *t*-tests, we also ran a Bayesian repeated measures ANOVA of all of the full sample focusing on the 2 (verbal and visuospatial WM load) × 2 (multiplication and subtraction) interaction. Comparison of model Bayes factors (BF_{10}) can be found in Tables A8 and A9. While the tables use the null model as a reference, it is more useful to compare Bayes factor between an additive model (WM task + arithmetic) and the interaction-included model (WM task + arithmetic + WM task × arithmetic). Comparing model fit between the two can be accomplished by taking the ratio of the Bayes factor of the additive model to the interaction-included model. The inverse of the ratio would provide a Bayes factor of the interaction alone compared to the null. The Bayesian ANOVA indicated anecdotal to moderate support for the additive model over the interaction-included model. The Bayesian ANOVA indicated that Bayes factors for additive model of WM task and arithmetic fit both reaction time and accuracy data better across the whole sample than with the additive and WM task × arithmetic operation interaction term included (RT BF_{10} ratio: 6.29; ACC BF_{10} ratio: 3.69). Higher BF_{10} ratios indicate greater support for the additive model over the interaction model.

As a preregistered robustness check, we estimated the same models for three subsamples of the data: easier secondary task blocks, more difficult secondary task blocks, and the first arithmetic block under cognitive load only. Similar patterns of results for both frequentist and Bayesian analyses can be found in our secondary analyses of the easier load, harder load, and first block conditions (Figures A2 and A3 in the Appendix and Tables S1, S2, S5, and S6 in the online Supplementary Materials); nearly all Bayesian estimates provide support for the null hypothesis. Only in the harder difficulty load condition[2] was there an effect on reaction time consistent with Hypothesis 1a, $F(1, 96) = 6.57$, $p < .05$, $\eta_p^2 = .07$. A post-hoc pairwise *t*-test of the hard load condition revealed a small but significant slowing in multiplication RT of 39 ms when under verbal load compared to visuospatial load, $t(96) = 2.67$, $p = .017$, $d = 0.16$, Holm-Bonferroni corrected.

2) The staircase procedure used during the first session of each experiment to estimate each participant's subjective 80th and 99th percentile threshold for their verbal and visuospatial cognitive loads provided reasonable estimates. On average, the 99th percentile (easy load) threshold for participants was 5.52 ($SD = 1.28$) for their verbal WM load and 6.52 ($SD = 0.95$) for their visuospatial WM load. For the 80th percentile (hard load), the threshold for participants' verbal WM load was 7.52 ($SD = 1.28$) and 8.52 ($SD = 0.95$) for their visuospatial WM load.

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

Again, in contrast to our hypothesis, in the full sample, subtraction performance was not significantly slower (Figure A2) nor was it was less accurate (Figure A3) under VSSP load compared to PL load. The ANOVA results shown in Table A4 yielded no significant difference in subtraction reaction time [RT: $F(1, 96) = 0.15$, $p = .70$, $\eta_p^2 = .002$]. Results shown in Table A5 yielded a significant difference in accuracy [ACC: $F(1, 96) = 6.31$, $p = .01$, $\eta_p^2 = .06$] between verbal and visuospatial dual-task load. However, this effect was in the opposite direction as predicted: Post-hoc pairwise *t*-test of the whole sample yielded a statistically significant decrease in subtraction accuracy of about 2 percentage points when under verbal load compared to visuospatial load [Whole: $t(96) = 2.57$, $p = .04$, $d = 0.19$, Holm-Bonferroni corrected]. Bayesian *t*-tests of WM load on subtraction RT and ACC found stronger evidence for the null hypothesis than for Hypothesis 1b such that there was no difference in subtraction RT by WM load type (RT BF_{01}: 8.29), but there was a difference in accuracy in favor of verbal load (ACC BF_{01}: 0.46). $BF_{01} > 1$ indicate more support for the null than the alternative while $0 \leq BF_{01} \leq 1$ indicate greater support for the alternative. Our Bayesian repeated measures ANOVA from the previous section included subtraction in the model, thus they can be applied here as well (also see Tables A8 and A9).

As a preregistered robustness check, we estimated the same models for the easy, hard, and first-arithmetic block under load subsamples of the data. Similar patterns of reaction time results for both frequentist and Bayesian analyses can be found in our secondary analyses of the easier load, harder load and first block conditions (Figures A2 and A3 in the Appendix and Tables S1, S2, S5, and S6 in the Supplementary Materials). In accuracy, we found a significant effect of WM load type within the easy load and first cognitive load block, $F(1, 96) = 7.44$ and 5.77, $p = .01$ and 0.02, $\eta_p^2 = .07$ and .06, respectively. However, this effect was consistent with what was found for the whole group, such that verbal load *lowered* accuracy more than visuospatial load [Easy: $t(96) = 2.73$, $p = .01$, $d = 0.23$, Holm-Bonferroni corrected; First: $t(96) = 2.40$, $p = .02$, $d = 0.22$, Holm-Bonferroni corrected], opposite of theoretical predictions.

Secondary Hypotheses:

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

To test Hypothesis 1c we included the single multiplication task condition into the 2-way ANOVA and performed pairwise *t*-test comparisons with Holm-Bonferroni corrections as needed. We did not find support for Hypothesis 1c: multiplication performance under no load was significantly faster (Figure A2) and more accurate (Figure A3) than both load conditions across most subsamples. There was a significant main effect of load vs. no load on multiplication reaction time for the whole sample, $F(1, 96) = 74.90$, $p < .001$, $\eta_p^2 = .44$. WM load yielded an average slowing of 143 ms or 18% ($d = 0.64$) in the whole sample. Mean comparisons and post-hoc pairwise *t*-test results are shown in Table A6. Reaction times under both verbal and visuospatial load were significantly slower than multiplication alone [RT: both $t(96) > 9.70$, $p < .001$, $d =$ (vs. Verbal: 0.68; vs. Visuospatial: 0.60)]. There was also a significant effect of load on multiplication accuracy for the whole sample, $F(1, 96) = 12.92$, $p < .001$, $\eta_p^2 = .12$, with accuracy being reduced by about 3% ($d = 0.30$). Mean comparisons and post-hoc pairwise *t*-test results are shown in Table A7. Accuracy comparisons were significant for the whole sample [ACC: both $t(96) > 3.99$, $p < .001$, $d =$ (vs. Verbal: 0.32; vs. Visuospatial: 0.28)]. We included the no load level into our Bayesian repeated measures ANOVA. Our Bayesian ANOVA indicated moderate to strong support for the additive model over the interaction-included model. Tables A10 and A11 of our Bayesian ANOVA indicated that even with the inclusion of single task arithmetic, the combination of WM task and arithmetic operation fit both reaction time and accuracy data better across the whole sample than with the inclusion of the WM task × arithmetic operation interaction term (RT BF_{10} ratio: 19.12; ACC BF_{10} ratio: 13.90). Higher BF_{10} ratios indicate greater support for the additive model over the interaction model.

For our preregistered robustness check, we estimated the same models for the easy, hard, and first-arithmetic block under load subsamples of the data. Our frequentist and Bayesian analyses for our subsample analyses yielded similar patterns of results to our whole sample analyses (Tables S3, S4, S7, and S8 in the Supplementary Materials). Only in the easier load condition was there no significant difference in accuracy between single multiplication and multiplication under visuospatial load, $p = .62$.

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

We found no support for Hypothesis 1d either. Subtraction performance under no load was significantly faster than either load condition (Figure A2) and more accurate than either load condition (Figure A3) across all subsamples. There

was a significant main effect of load vs. no load on subtraction reaction time for the whole sample, $F(1, 96) = 62.28$, $p < .001$, $\eta_p^2 = .39$. WM load yielded an average slowing of 139 ms or 19% ($d = 0.59$) in the whole sample. Mean comparisons and post-hoc pairwise *t*-test results are shown in Table A6. Reaction times under both verbal and visuospatial load were significantly slower than subtraction alone [RT: both $t(96) > 9.76$, $p < .001$, $d =$ (vs. Verbal: 0.61; vs. Visuospatial: 0.57)]. There was a significant effect of WM load on accuracy as well [ACC: $F(1, 96) = 13.54$, $p < .001$, $\eta_p^2 = .12$] with about a 3% ($d = 0.31$) reduction in accuracy under load in the whole sample. Mean comparisons and post-hoc pairwise *t*-test results are shown in Table A7. Subtraction accuracy was weaker under verbal load compared to no load, $t(96) = 5.23$, $p < .001$, $d = 0.37$, but not between no load and visuospatial load ($p = .07$). Our Bayesian repeated measures ANOVA results are the same as reported for Hypothesis 1c.

For our preregistered robustness check, we estimated the same models for the easy, hard, and first-arithmetic block under load subsamples of the data. Both frequentist and Bayesian analyses for the subsample analyses yielded similar patterns of results to our whole sample analyses (Tables S3, S4, S7, and S8 in the Supplementary Materials).

Comparing US- vs. Chinese-Educated Participants: Hypotheses 2a-2d

To test whether the differential influence of working memory depends on where students received their primary math education, we computed a 2 (country; US- vs. Chinese-educated) × 2 (WM load) × 2 (arithmetic) ANOVA in order to test whether the differential impact of WM load type on arithmetic operation is dependent on where participants received the majority of their math education. The 3-way ANOVA did not yield a significant main effect for country, $F(1, 91) = 0.56$, $p = .46$, $\eta_p^2 = .01$, but it did yield a significant main effect for arithmetic operation, $F(1, 91) = 8.31$, $p = .005$, $\eta_p^2 = .08$. Furthermore, the ANOVA did not yield a significant 3-way interaction for country × WM task × arithmetic, $F(1, 91) = 1.57$, $p = .21$, $\eta_p^2 = .02$, nor 2-way interactions for WM task × country, $F(1,91) = 0.27$, $p = .60$, $\eta_p^2 = .003$, or WM task × arithmetic, $F(1, 91) = 2.73$, $p = .10$, $\eta_p^2 = .03$. However, there was a significant 2-way interaction for and country × arithmetic, $F(1, 91) = 4.25$, $p = .04$, $\eta_p^2 = .05$. Post-hoc pairwise *t*-test comparisons revealed that the US-educated participants were generally slower in multiplication than in subtraction by about 100 ms [$t(91) = 5.08$, $p < .001$, $d = 0.43$, Holm-Bonferroni corrected] while no such difference in reaction times were present in the Chinese-educated participants.

In accuracy, there were no significant effects for country, $F(1, 91) = 3.30$, $p = .07$, $\eta_p^2 = .04$, or any 3-way or 2-way interactions. The 3-way ANOVA yielded only main effects for WM task, $F(1, 91) = 6.38$, $p = .01$, $\eta_p^2 = .07$, and arithmetic, $F(1, 91) = 4.29$, $p = .04$, $\eta_p^2 = .05$. Taken together, these findings provide some support for the validity of two of the sources of variation in our population: First, the WM load manipulations were sufficiently difficult to impair arithmetic performance. Second, Chinese-educated students showed a different pattern of performance on arithmetic tasks, being approximately equally fast and accurate at multiplication and subtraction, relative to the US-educated participants, which were consistently faster and more accurate at subtraction than multiplication.

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

Following the lack of a 3-way interaction, we examined the Chinese-educated subgroup directly. Overall, we did not find evidence to support Hypothesis 2a. While there appeared to be a moderate effect of verbal vs. visuospatial load on multiplication reaction times (see Table A1, row 5, column 5), this effect was not statistically significant, $d = 0.28$, $p = .51$. Our ANOVA results in Tables A4 and A5 also suggest that WM load type did not differentially impact multiplication performance [RT: $F(1, 21) = 1.69$, $p = .21$, $\eta_p^2 = .07$; ACC: $F(1, 21) = 3.59$, $p = .07$, $\eta_p^2 = .15$]. Bayesian pairwise *t*-tests for reaction times and accuracy produced $BF_{01} = 2.14$ and 0.99, respectively, suggesting anecdotal evidence in favor of the null hypothesis. $BF_{01} > 1$ indicate more support for the null than the alternative while BF_{01} approaching 1 suggest no evidence for either null or alternative. Bayesian repeated measures ANOVA models in Tables A8 and A9 indicated a better fit for the additive (WM load type + arithmetic operation) model over the additive + interaction model as well (RT BF_{10} ratio = 2.48; ACC BF_{10} ratio = 3.27). Higher BF_{10} ratios indicate greater support for the additive model over the interaction model.

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

Overall, we did not find evidence to support Hypothesis 2b. The effect of visuospatial load on subtraction reaction times had a much smaller effect size than in Lee and Kang (2002), (see Table A1, row 5, column 6), but was not

statistically significant either, d = -0.09, p = .98. Our ANOVA results in Tables A4 and A5 also suggest that WM load type did not differentially impact subtraction performance [RT: $F(1, 21) = 0.17$, $p = .67$, $\eta_p^2 = .01$; ACC: $F(1, 21) = 3.41$, $p = .08$, $\eta_p^2 = .14$]. Bayesian pairwise t-tests for reaction time and accuracy produced BF_{01} = 4.15 and 1.06, suggesting anecdotal evidence in favor of the null hypothesis. $BF_{01} > 1$ indicate more support for the null over the alternative. Similarly, our Bayesian repeated measures ANOVA models in Tables A8 and A9 found better fit for the additive (WM load type + arithmetic operation) model over the additive + interaction model as well (RT BF_{10} ratio = 2; ACC BF_{10} ratio = 3.27).

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

We did not find evidence to support Hypothesis 2c. Multiplication performance under no load was significantly faster than both load conditions (Figures A2 and A3). There was a significant main effect of load on multiplication reaction time but not accuracy for the Chinese-educated participants [RT: $F(2, 42) = 13.41$, $p < .001$, $\eta_p^2 = .39$; ACC: $F(2, 42) = 2.84$, $p = .07$, $\eta_p^2 = .12$]. Mean comparisons and post-hoc pairwise t-test results are shown in Tables A6 and A7. Overall, multiplication reaction time was impacted by both load types (RT: vs. Verbal: d = 0.79, BF_{10} = 417; vs. Visuospatial: d = 0.67, BF_{10} = 31). $BF_{10} > 1$ indicate greater support for the alternative that there was a difference in performance between no load and either secondary task load. The Bayesian repeated measures ANOVA reported similar patterns of results as those from Hypothesis 2a and 2b with the additive (WM load type + arithmetic operation) model fitting the data better than the additive + interaction model (RT BF_{10} ratio = 4.71; ACC BF_{10} ratio = 4.4; Tables A10 and A11).

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

We did not find evidence to support Hypothesis 2d. Subtraction performance under no load was significantly faster than both load conditions (Figures A2 and A3). There was a significant main effect of load on multiplication reaction time but not accuracy for the Chinese-educated sample [RT: $F(2, 42) = 5.02$, $p = .01$, $\eta_p^2 = .19$; ACC: $F(2, 42) = 0.97$, $p = .39$, $\eta_p^2 = .04$]. Mean comparisons and post-hoc pairwise t-test results are shown in Tables A6 and A7. Overall, subtraction reaction time was impacted by both load types (RT: vs. Verbal: d = 0.36, BF_{10} = 1.36; vs. Visuospatial: d = 0.41, BF_{10} = 7.19). $BF_{10} > 1$ indicate greater support for the alternative that there was a difference in performance between no load and either secondary task load. The Bayesian repeated measures ANOVA from Tables A10 and A11 are the same as those reported for Hypothesis 2c.

Discussion

In this registered report, we tested several pre-registered predictions based on previous findings from the dual-task literature with respect to the differential effects of secondary WM task load on arithmetic performance. That is, we tested whether verbal secondary tasks reduce multiplication performance but not subtraction performance, whether visuospatial secondary tasks reduce subtraction performance but not multiplication performance, if these differential effects can be observed relative to each other or a no load control. These predictions have implications for theories of mathematical cognition and working memory, along with dual-task performance specifically. Building upon previous work in the field, we identified potential moderators that could explain contradictory findings from the literature - specifically, secondary task difficulty and having learned mathematics primarily in China – and tested whether hypothesized effects emerge under these conditions.

Consistent with all previous literature, we found arithmetic performance to be generally slower and less accurate under cognitive load. However, contrary to our pre-registered predictions, we found no evidence for the moderating effect of secondary WM task load on arithmetic operations performance across the whole sample (Hypotheses 1a-1d) or any of our preregistered subgroup analyses, nor did we find evidence for these effects in the Chinese-educated participants (Hypotheses 2a-2d). In our follow-up Bayesian analyses, our results generally provided support for the null hypothesis that there was no moderation by secondary WM load on arithmetic operation over the additive effects of secondary WM task and arithmetic operation. However, the majority of Bayes factors suggested only anecdotal evidence for the null over the alternative, suggesting that these differential effects could still be real but that our current

experiment was unable to find sufficient evidence otherwise. At best, we found anecdotal evidence of a verbal WM load effect on subtraction accuracy across some of our subsample analyses; however, the direction of this effect was the opposite of what was predicted by previous work. In sum, we did not find any evidence for the large strong crossover interaction reported by Lee and Kang (2002).

Interactions between secondary task types and arithmetic play a prominent role in the dual-task literature. These interactions have been interpreted as providing evidence that domain-specific pathways, such as verbal or visuospatial pathways have differential effects on numerical cognition (e.g., Ashcraft, 1992; Dehaene, 1992; Dehaene & Cohen, 1995, 1997; see Chen & Bailey, 2021, for review). However, we argue that there are strong theoretical and empirical reasons to reexamine the robustness of these interactions. Several theoretical accounts of working memory argue against multiple domain-specific influences in favor of a more centralized executive processing system (Barrouillet & Camos, 2001; Cowan, 1999; Engle, 2002; Oberauer, 2009). Further, two other studies since Lee and Kang (2002) were also unable to replicate they key crossover interaction reported in the original paper (Cavdaroglu & Knops, 2016; Imbo & LeFevre, 2010). Imbo and LeFevre (2010) reported a differential effect in accuracy among Chinese students, such that there were more multiplication errors under verbal load compared to visuospatial load, but no differential effects of visuospatial vs. verbal load on subtraction performance. Both prior studies used a mix of within and between subject factors in their design. In comparison, our fully within-subjects study did not find any differential effects of WM load in our Chinese-educated participants nor across difficulty levels. To our knowledge, Lee and Kang (2002) remain the only study to have reported this crossover effect. Given the small sample size of the previous study (n = 10) and lack of subsequent replication, we propose that the field should consider the possibility that such crosstalk effects may be idiosyncratic to particular combinations of primary and secondary tasks and/or the particular population – irrespective of where they received their primary math education. Thus, crosstalk effects may be difficult to predict a priori. This view could certainly be replaced by a theory that can 1) account for when crossover interactions occur or not in the previous literature, and 2) make predictions about replicable effects in future work.

Additionally, our results conflict with parallel processing models of dual-task theory that attribute differences in dual-task performance to the amount of overlap in cognitive resources between two tasks (Navon & Miller, 2002; Tombu & Jolicœur, 2003; Wickens, 2008). While Pashler (1994) notes that crossover effects are still possible in the absence of parallel processing and may explain similar effects seen in sequential processing theories, the reasons for this may be specific to the combination of primary and secondary tasks and difficult to predict a priori (for another review, see Fischer & Plessow, 2015). For example, Hubber et al. (2014) had participants complete addition tasks alone and with a visuospatial task (i.e., remembering patterns of colored blocks) and initially found that visuospatial memory moderated the types of strategies used in addition. The visuospatial task included an n-back component in which they had to remember if the target block was the same as the one presented before the previous block, so a follow up experiment was done in which a more static visuospatial task was used (i.e., without n-back component) and a separate central executive task was used (i.e., random letter generation). The follow-up found no difference in arithmetic performance or strategy selection between the single task condition and the dual-task with visuospatial load, but a major difference between the single task and dual-task with central executive load, suggesting that evidence of a parallel processing effect was confounded by the complexity of the secondary task. If dual-task performance relies on sequential processing, the current study still provides evidence for the effect of working memory on arithmetic performance, but the cost of performance caused by a potential cognitive bottleneck is likely more domain-general in nature than what is commonly assumed in the literature (for review, see Doherty et al., 2019).

To conclude, the current study investigated the differential effect of WM task loads (verbal and visuospatial) on arithmetic operations (multiplication and subtraction). Consistent with prior meta-analytic work on correlations between WM tasks and arithmetic performance and the dual-task literature on WM and arithmetic performance, the current study found consistent effects on arithmetic performance when under load of more complex secondary tasks, but no clear pattern for domain-specific interference. Despite investigating whether the crossover effect would emerge under conditions previously hypothesized to moderate the effect (difficulty and the system in which participants were educated), we did not find evidence for the predicted interaction in any of our analyses. Although multiplication and subtraction seemed to operate exclusively through verbal and visuospatial pathways, respectively, in the original study, this interaction has not been subsequently observed. We interpret these findings as evidence for a more domain-general

pathway for WM secondary tasks' influence on numerical cognition, although we encourage future work that continues to carefully consider how theories of working memory and dual-task performance could explain previous domain-specific effects within numerical cognition.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Ethics Statement: Research was carried out in accordance with the ethical principles and standards of the Institutional Review Board at the University of California, Irvine.

Data Availability: For this article, a data set is freely available (Chen, Bailey, & Jaeggi, 2022a).

Supplementary Materials

The Supplementary Materials contain the following items (for access see Index of Supplementary Materials below).:

- Preregistration protocol
- Research data and codebook
- R code to organize data for analyses in JASP
- Arithmetic tasks, additional analyses, and documentation of deviations from original pre-registration
- Protocol for dual-task experiment

Index of Supplementary Materials

Chen, E. H., Bailey, D. H., & Jaeggi, S. M. (2021). *Supplementary materials to "No clear support for differential influences of visuospatial and phonological resources on mental arithmetic: A Registered Report"* [Preregistration protocol]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.4871

Chen, E. H., Bailey, D. H., & Jaeggi, S. M. (2022a). *Supplementary materials to "No clear support for differential influences of visuospatial and phonological resources on mental arithmetic: A Registered Report"* [Research data and codebook]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.6882

Chen, E. H., Bailey, D. H., & Jaeggi, S. M. (2022b). *Supplementary materials to "No clear support for differential influences of visuospatial and phonological resources on mental arithmetic: A Registered Report"* [R code to organize data for analyses in JASP]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.6880

Chen, E. H., Bailey, D. H., & Jaeggi, S. M. (2022c). *Supplementary materials to "No clear support for differential influences of visuospatial and phonological resources on mental arithmetic: A Registered Report"* [Arithmetic tasks, additional analyses, and documentation of deviations from original pre-registration]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.6881

Chen, E. H., Bailey, D. H., & Jaeggi, S. M. (2022d). *Supplementary materials to "No clear support for differential influences of visuospatial and phonological resources on mental arithmetic: A Registered Report"* [Protocol for dual-task experiment]. PsychOpen GOLD. https://doi.org/10.23668/psycharchives.7049

References

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

Barrouillet, P., & Camos, V. (2001). Developmental increase in working memory span: Resource sharing or temporal decay? *Journal of Memory and Language, 45*(1), 1–20. https://doi.org/10.1006/jmla.2001.2767

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. https://doi.org/10.1007/s00426-015-0667-8

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), 220–233. https://doi.org/10.1037/xlm0000822

Cowan, N. (1999). An embedded-processes model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62–101). Cambridge University Press. https://doi.org/10.1017/CBO9781139174909.006

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

Dehaene, S., & Cohen, L. (1995). Towards an anatomical and functional model of number processing. *Mathematical Cognition, 1*(1), 83–120.

Dehaene, S., & Cohen, L. (1997). Cerebral pathways for calculation: Double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex, 33*(2), 219–250. https://doi.org/10.1016/S0010-9452(08)70002-9

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology, 20*(3-6), 487–506. https://doi.org/10.1080/02643290244000239

Doherty, J. M., Belletier, C., Rhodes, S., Graham, A. J., Barrouillet, P., Camos, V., Cowan, N., Naveh-Benjamin, M., & Logie, R. H. (2019). Dual-task costs in working memory: An adversarial collaboration. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(9), 1529–1551. https://doi.org/10.1037/xlm0000668

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19–23. https://doi.org/10.1111/1467-8721.00160

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. https://doi.org/10.3758/BRM.41.4.1149

Fischer, R., & Plessow, F. (2015). Efficient multitasking: Parallel versus serial processing of multiple tasks. *Frontiers in Psychology, 6*, Article 1366. https://doi.org/10.3389/fpsyg.2015.01366

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*(2), 65–70.

Hubber, P. J., Gilmore, C., & Cragg, L. (2014). The roles of the central executive and visuospatial storage in mental arithmetic: A comparison across strategies. *Quarterly Journal of Experimental Psychology, 67*(5), 936–954. https://doi.org/10.1080/17470218.2013.838590

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. https://doi.org/10.3758/MC.38.2.176

JASP Team. (2020). *JASP* (Version 0.16.1) [Computer software].

Jeffreys, H. (1961). *Theory of probability* (3rd ed.). New York, NY, USA: Oxford University Press.

Kawashima, R., Taira, M., Okita, K., Inoue, K., Tajima, N., Yoshida, H., Sasaki, T., Sugiura, M., Watanabe, J., & Fukuda, H. (2004). A functional MRI study of simple arithmetic: A comparison between children and adults. *Brain Research: Cognitive Brain Research, 18*(3), 227–233. https://doi.org/10.1016/j.cogbrainres.2003.10.009

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. https://doi.org/10.1002/1531-8249(200010)48:4<657::AID-ANA13>3.0.CO;2-K

Lee, K. M., & Kang, S. Y. (2002). Arithmetic operation and working memory: Differential suppression in dual tasks. *Cognition, 83*(3), B63–B68. https://doi.org/10.1016/S0010-0277(02)00010-0

Morey, R. D., & Rouder, J. N. (2011). Bayes factor approaches for testing interval null hypotheses. *Psychological Methods, 16*(4), 406–419. https://doi.org/10.1037/a0024377

Navon, D., & Miller, J. (1987). Role of outcome conflict in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 13*(3), 435–448. https://doi.org/10.1037/0096-1523.13.3.435

Navon, D., & Miller, J. (2002). Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology, 44*(3), 193–251. https://doi.org/10.1006/cogp.2001.0767

Oberauer, K. (2009). Design for a working memory. *Psychology of Learning and Motivation, 51*, 45–100. https://doi.org/10.1016/S0079-7421(09)51002-X

Pashler, H. (1994). Graded capacity-sharing in dual-task interference? *Journal of Experimental Psychology: Human Perception and Performance, 20*(2), 330–342. https://doi.org/10.1037/0096-1523.20.2.330

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods, 51*, 195–203. https://doi.org/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. https://doi.org/10.1002/hbm.21159

RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA, USA. http://www.rstudio.com

Tombu, M., & Jolicœur, P. (2003). A central capacity sharing model of dual-task performance. *Journal of Experimental Psychology: Human Perception and Performance, 29*(1), 3–18. https://doi.org/10.1037/0096-1523.29.1.3

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors, 50*(3), 449–455. https://doi.org/10.1518/001872008X288394

Wickham, H., Averick, M., Bryan, J., Winston, C., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., . . . Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software, 4*(43), Article 1686. https://doi.org/10.21105/joss.01686

Appendix

Table A1

Studies That Tested Arithmetic Operation × WM Load Type Interaction

Author	Sample size	WM tasks	Arithmetic tasks	Multiplication effect (PL vs. VSSP) *d*	Subtraction effect (PL vs. VSSP) *d*	PL vs. VSSP in Multiplication; Subtraction (ms)
Lee, K. M., & Kang, S. Y. (2002)	10	Repeat nonword string (PL), Matching abstract shapes and location (VSSP)	exact subtraction, exact multiplication	2.42	-3.31	1170 vs. 996 993 vs. 1271
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	0.04	0.04	5103 vs. 5018 4823 vs. 4738
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	-0.02	0.07	3015 vs. 3038 3068 vs. 2998
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	0.00	1015 vs. 989 864 vs. 863
Chen, E. H., Jaeggi, S. M., & Bailey, D. H. – Chinese-educated sample	22	Letter span (PL), 5×5 grid location task (VSSP)	2AFC multiplication (one × one; two × one digit), 2 AFC subtraction (two – one digit)	0.28	-0.09	883 vs. 841 840 vs. 851
Chen, E. H., Jaeggi, S. M., & Bailey, D. H. – other-educated sample	71	Letter span (PL), 5×5 grid location task (VSSP)	2AFC multiplication (one × one; two × one digit), 2 AFC subtraction (two – one digit)	0.05	-0.02	946 vs. 939 841 vs. 842

Note. Cohen's *d* were calculated for the columns 5 and 6. Cohen's *d* represents effect size between multiplication and subtraction RT performance under PL or VSSP load, respectively. PL – Phonological/Verbal; VSSP – Visuospatial. Reaction times rounded to nearest ms.

Table A2

Means and Standard Deviations for Reaction Time and Accuracy as a Function of WM Task × Arithmetic Operation

WM task	Arithmetic	RT Mean	*SD*	ACC Mean	*SD*	*N*
No load	Multiplication	787	222	93%	11%	97
	Subtraction	713	213	95%	10%	97
Verbal	Multiplication	938	221	90%	11%	97
	Subtraction	850	234	91%	8%	97
Visuospatial	Multiplication	923	231	90%	9%	97
	Subtraction	854	276	93%	8%	97

Note. WM = working memory. Reaction times (RT) in nearest ms and accuracy (ACC) in nearest percentage.

Table A3

Demographic Information

Variable	*N* = 97	*M* (*SD*)	%
Gender			
Male	33		66
Female	64		34
Age		20.1 (1.3)	
Country of primary math education			
US	71		73.2
China	22		22.7
Other	4		4.1
Math grade compared to peer			
A	35		36.08
B	49		50.52
C	12		12.37
D	1		1.03
F	0		0
Abacus use			
Never Taught	69		71.13
Never Used	16		16.49
Rarely	11		11.34
Sometimes	1		1.03

Table A4

Planned ANOVA on Arithmetic Reaction Time by Model Specification

	Whole sample				Chinese-educated			
Factor	***F***	***df***	***p***	η_p^2	***F***	***df***	***p***	η_p^2
PL vs. VSSP × Multiplication	1.20	(1, 96)	.28	.01	1.69	(1, 21)	.21	.07
PL vs. VSSP × Subtraction	.15	(1, 96)	.70	.002	.17	(1, 21)	.67	.01

Note. PL – Phonological/Verbal load; VSSP – Visuospatial load. Whole sample – no restriction on participants, Chinese-educated – only participants that reported their primary math education came from China.

Table A5

Planned ANOVA on Arithmetic Accuracy by Model Specification

	Whole sample				Chinese-educated			
Factor	***F***	***df***	***p***	η_p^2	***F***	***df***	***p***	η_p^2
PL vs. VSSP × Multiplication	.49	(1, 96)	.49	.01	3.59	(1, 21)	.07	.15
PL vs. VSSP × Subtraction	6.31*	(1, 96)	.01	.06	3.41	(1, 21)	.08	.14

Note. PL – Phonological/Verbal load; VSSP – Visuospatial load. Whole sample – no restriction on participants, Chinese-educated – only participants that reported their primary math education came from China.
*$p < .05$.

Table A6

Comparison of Reaction Time With no Load Arithmetic

Model	**Task**	**Mean difference**	***SE***	***t***	***df***	p_{holm}
Multiplication						
Whole	Verbal	151	13.98	10.77	96	< .001
	Visuospatial	136	13.98	9.71	96	< .001
Chinese	Verbal	160	34.45	4.65	21	< .001
	Visuospatial	118	34.45	3.42	21	.014
Subtraction						
Whole	Verbal	137	13.98	9.77	96	< .001
	Visuospatial	141	13.98	10.12	96	< .001
Chinese	Verbal	94	34.45	2.74	21	< .001
	Visuospatial	106	34.45	3.07	21	< .001

Note. Mean difference is reaction time in nearest ms. p_{holm} – p value after Holm-Bonferroni correction.

Table A7

Comparison of Accuracy With no Load Arithmetic

Model	**Task**	**Mean difference**	***SE***	***t***	***df***	p_{holm}
Multiplication						
Whole	Verbal	-3%	1	-4.68	96	< .001
	Visuospatial	-3%	1	-3.99	96	< .001
Chinese	Verbal	-4%	2	-2.67	21	.14
	Visuospatial	-2%	2	-1.39	21	1
Subtraction						
Whole	Verbal	-4%	1	-5.23	96	< .001
	Visuospatial	-2%	1	-2.64	96	.07
Chinese	Verbal	-2%	2	-1.22	21	1
	Visuospatial	0%	2	0.22	21	1

Note. Mean difference is accuracy in nearest percentage. p_{holm} – p value after Holm-Bonferroni correction.

Table A8

Bayesian Model Comparisons of 2×2 ANOVA on Reaction Time

	Whole		Chinese	
Models	**BF_{10}**	**error %**	**BF_{10}**	**error %**
Arithmetic	3.80e +6	0.91	0.25	1.33
WM task + Arithmetic	4.50e +5	1.78	0.06	1.42
WM task + Arithmetic + WM task × Arithmetic	9.60e +4	4.67	0.03	2.38
WM task	0.12	0.79	0.26	1.92

Note. All models include subject. Null model is used as reference.

Table A9

Bayesian Model Comparisons of 2×2 ANOVA on Accuracy

	Whole		Chinese	
Models	**BF_{10}**	**error %**	**BF_{10}**	**error %**
Arithmetic	159.70	1.11	0.31	0.85
WM task + Arithmetic	106.89	3.19	0.49	1.73
WM task + Arithmetic + WM task × Arithmetic	28.05	2.00	0.15	2.86
WM task	0.59	0.90	1.58	1.68

Note. All models include subject. Null model is used as reference.

Table A10

Bayesian Model Comparisons of 3×2 ANOVA on Reaction Time

	Whole		Chinese	
Models	**BF_{10}**	**error %**	**BF_{10}**	**error %**
WM task	4.0e +26	0.87	808.66	0.81
Arithmetic	4.19e +7	0.73	0.19	2.02
WM task + Arithmetic	1.3e +37	1.02	153.86	2.02
WM task + Arithmetic + WM task × Arithmetic	6.8e +35	3.51	32.60	3.00

Note. All models include subject. Null model is used as reference.

Table A11

Bayesian Model Comparisons of 3×2 ANOVA on Accuracy

	Whole		Chinese	
Models	**BF_{10}**	**error %**	**BF_{10}**	**error %**
WM task	1.61e +5	0.92	1.21	0.86
Arithmetic	1868.54	1.21	0.18	1.26
WM task + Arithmetic	6.13e +8	2.00	0.22	1.34
WM task + Arithmetic + WM task × Arithmetic	4.41e +7	0.90	0.05	2.12

Note. All models include subject. Null model is used as reference.

Figure A1

Single Multiplication Task (A & C). Dual-Task Multiplication With Phonological Letter WM Load (B) and Visuospatial WM Load (D)

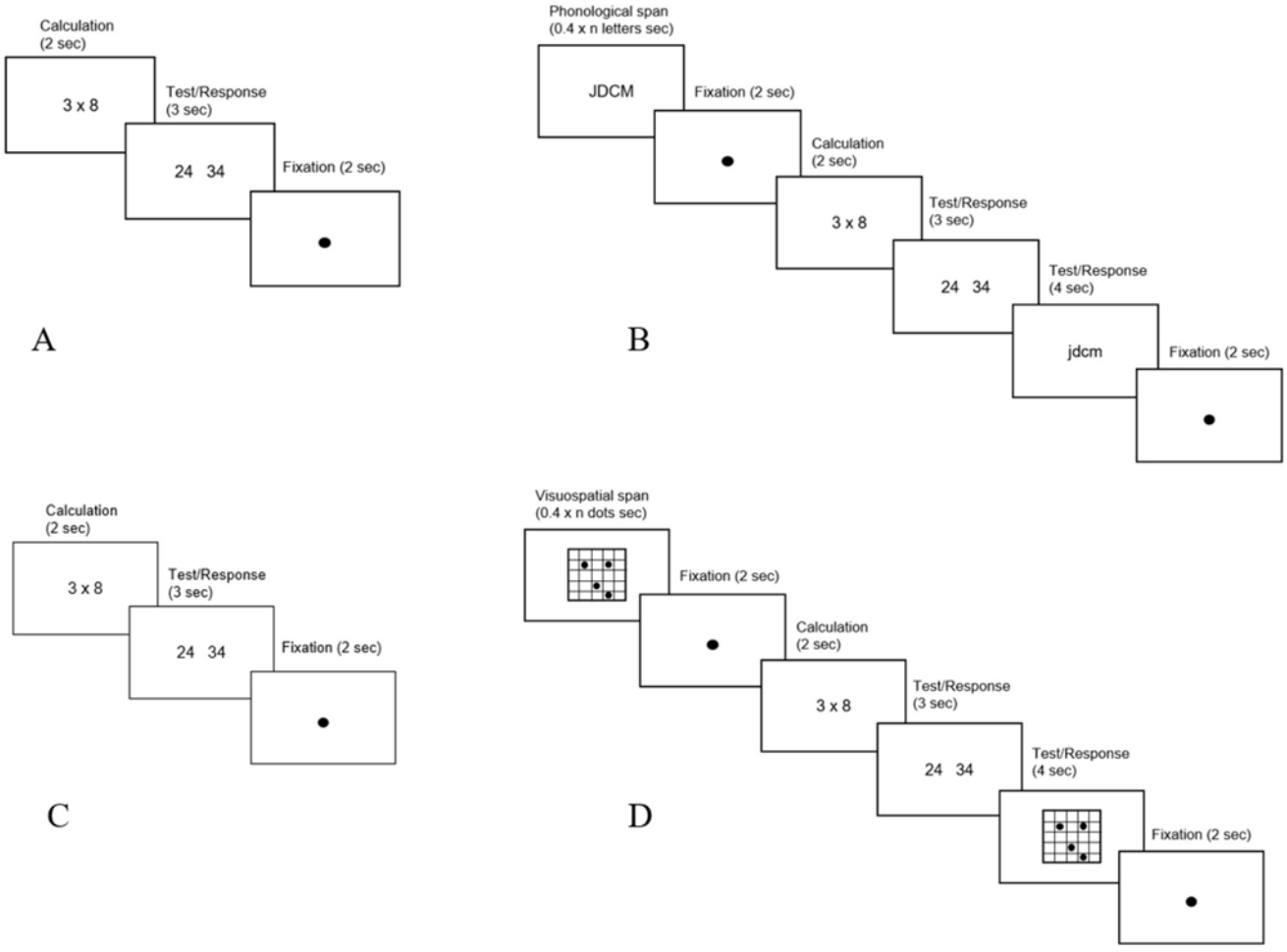
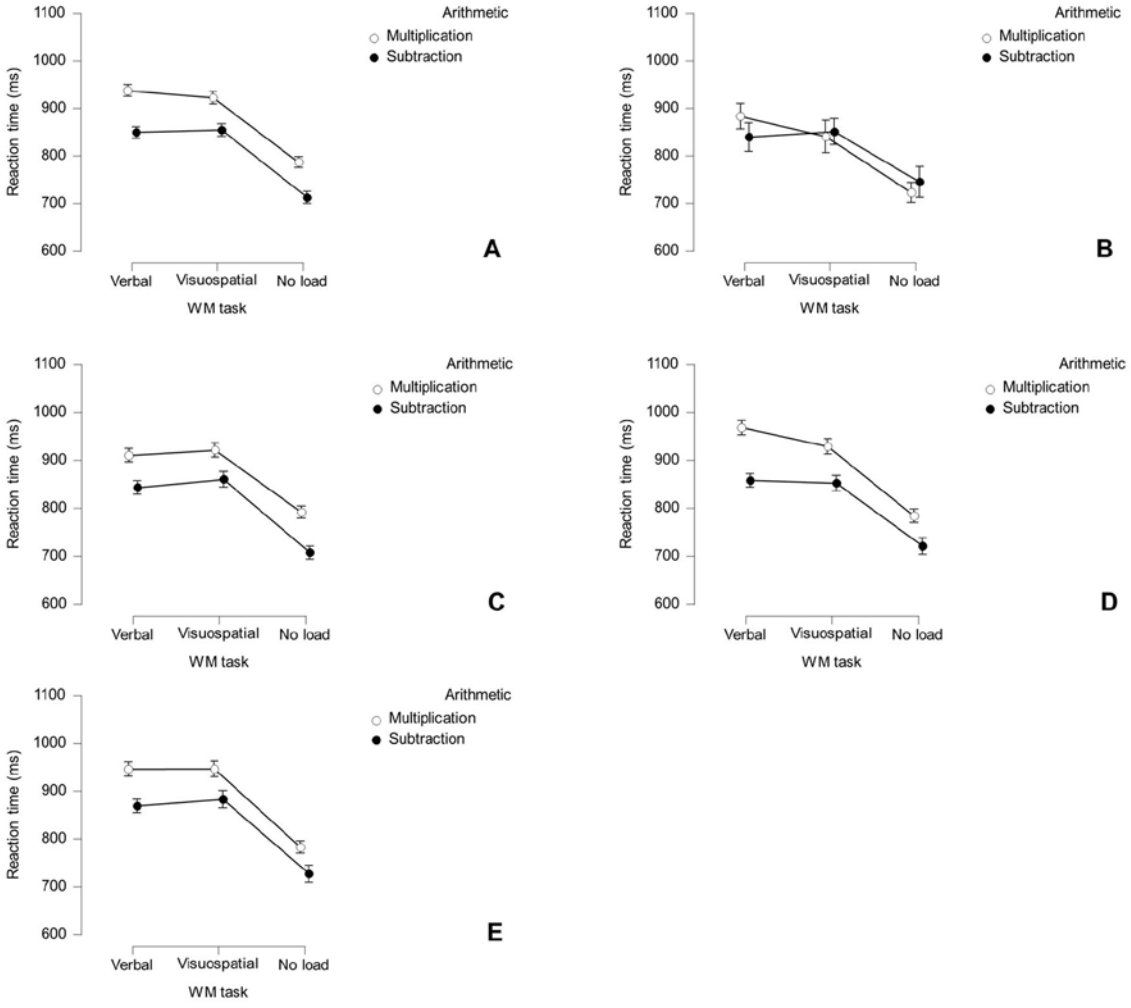


Figure A2

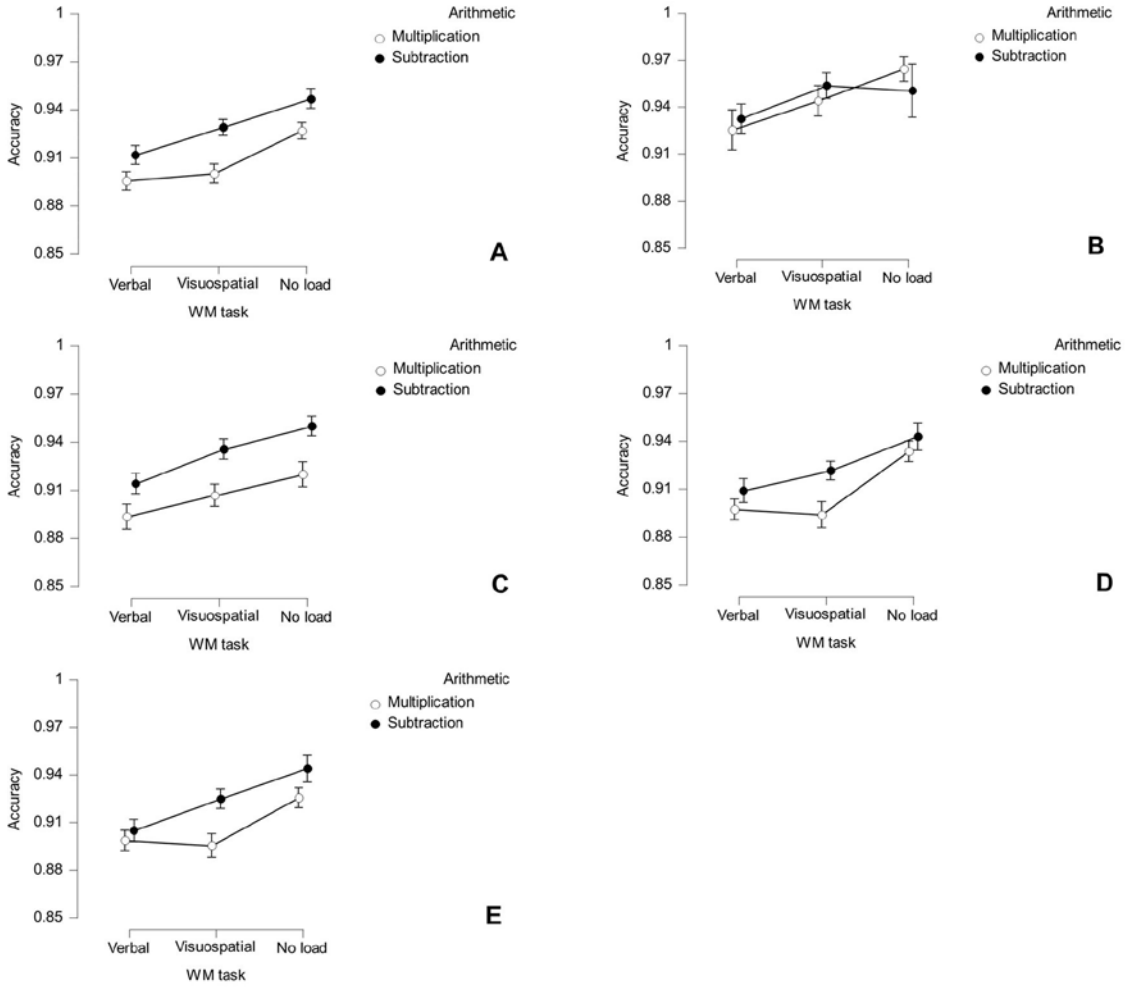
Comparison of Dual-Task Effects in Reaction Times Across Subsample Analyses



Note. A = Whole group; B = Received majority of math education in China; C = Easy load conditions only; D = Hard load conditions only; E = First under cognitive load conditions only. Error bars represent standard errors of the mean.

Figure A3

Comparison of Dual-Task Effects in Accuracy Across Subsample Analyses



Note. A = Whole group; B = Received majority of math education in China; C = Easy load conditions only; D = Hard load conditions only; E = First under cognitive load conditions only. Error bars represent standard errors of the mean.

Journal of Numerical Cognition (JNC) is an official journal of the Mathematical Cognition and Learning Society (MCLS).

leibniz-psychology.org

PsychOpen GOLD is a publishing service by Leibniz Institute for Psychology (ZPID), Germany.