Verification of multiplication facts: An investigation using retrospective protocols

STEPHEN G. ROMERO
University of Colorado at Boulder

TIMOTHY C. RICKARD
University of California, San Diego

LYLE E. BOURNE JR.
University of Colorado at Boulder

Retrospective verbal protocols collected throughout participants' performance of a multiplication verification task (e.g., "7 × 3 = 28, true or false?") documented a number of different strategies and changes in strategy use across different problem categories used for this common experimental task. Correct answer retrieval and comparison to the candidate answer was the modal but not the only strategy reported. Experiment 1 results supported the use of a calculation algorithm on some trials and the use of the difference between the candidate and correct answers (i.e., split) on others. Experiment 2 clearly demonstrated that participants sometimes bypassed retrieval by relying on the split information. Implications for mental arithmetic theories and the general efficacy of retrospective protocols are discussed.

Data relevant to theories of mental arithmetic come from both verification and production tasks. In a production task, participants are given a problem and asked to write, type, or say the correct answer (e.g., "3 × 4 = ?"). In a verification task, participants are given a problem and a candidate answer and are asked to indicate whether the candidate answer is true or false (e.g., "3 × 4 = 14, true or false?"). Although the use of both of these tasks has been helpful in elucidating the underlying cognitive processing (see Ashcraft, 1992, 1995, for an extensive review), specific questions regarding the relationship between the tasks still persist. These questions are of both methodological and theoretical interest. For example, the verification task has been viewed as a four-stage process of encoding, memory retrieval, comparison to the presented answer, and response execution (see Campbell, 1987b). In contrast, production has been thought to entail only encoding, memory retrieval, and response execution. It follows from this reasoning that the only theoretical dif-

AMERICAN JOURNAL OF PSYCHOLOGY
Spring 2006, Vol. 119, No. 1, pp. 87–120

ference between the tasks is the comparison process (i.e., verification is simply production plus comparison). This framework has provided the basis for development of many cognitive models of mental arithmetic in general and of verification specifically, and so a finding that there are processes used in verification different from those used in production would necessitate a reformulation of these models.

Arithmetic verification

Generally, verification studies show that participants reject incorrect answers more slowly than they accept correct answers and that incorrect answers that are related to the multiplication table of one of the operands are rejected more slowly than those that are not (i.e., relatedness effects). Moreover, a greater numerical split between presented false answers and the true answer results in faster responses. Using the framework outlined earlier, it has been suggested that the split and relatedness properties of the presented answer affect the memory retrieval or production processing stage (Stazyk, Ashcraft, & Hamann, 1982; see also Campbell, 1987b, 1991; Zbrodoff & Logan, 1990). In a model by Campbell (1987a, 1987b), the answer relatedness effects were explained as a product of differential priming for correct and incorrect answers. For true problems, the given answer primes the correct answer, facilitating retrieval. For false problems, associative priming from the given answer causes interference that slows down the retrieval process. Magnification of this interference for table-related problems accounts for differences in response times (RTs) between the two kinds of incorrect answers. Although this model was the first to account for the interfering effects of the presented answer in verification, it does not differ from the earlier theory developed by Ashcraft and Battaglia (1978) or from similar models proposed by Ashcraft (1982, 1987) in assuming that verification consists of production plus comparison.

In contrast to these production plus comparison analyses of verification, Zbrodoff and Logan (1990) speculated that production and verification tasks reflect very different underlying retrieval processes. Zbrodoff and Logan suggested that in verification formats, participants compare the equation with memory of earlier instances of the problem–answer combination as a whole and use this comparison to evaluate whether problem statement is true or false. This theory therefore assumes that participants do not retrieve a correct answer in the verification format but only match the problem statement as a whole (i.e., retrieve and match the entire problem representation).

Zbrodoff and Logan (1990) explained the differences in RTs for table-related and table-unrelated problems that they and Campbell (1987a, 1987b) observed as resulting from differential resonance, or relative strength of the equations in memory, indexed by the frequency of expo-

sure to each equation. Resonance is stronger for true problems than false problems because equations are seen more often in all contexts with the correct answer. Similarly, RT differences between the different types of false problems would result from stronger resonance for table-related than for table-unrelated incorrect answers. Evidence supporting this idea was presented from experiments manipulating the delay between the onset of the problem arguments (i.e., "4 × 7") and the onset of the candidate answer (i.e., "24"). Zbrodoff and Logan (1990) argued that that if verification is production plus comparison, delay between the arguments and answer should affect only the production processes of computation or retrieval and should not affect comparison processes. The persistence of problem difficulty effects at long delays and a decrease in split effects across delays was used as evidence that verification does not exclusively involve production plus comparison. Zbrodoff and Logan (1990) were careful to point out, however, that from their data they could not determine whether resonance was used on every trial in verification or as an occasional strategy used in conjunction with other strategies including production plus comparison.

Campbell and Tarling (1996) presented results that also seem to support the existence of resonance or familiarity processes in verification. Taking a cue from research on implicit and explicit forms of memory, they argued that production is primarily retrieval based (i.e., explicit), whereas verification is primarily familiarity based (i.e., implicit). In the Campbell and Tarling study, production and verification trials were alternated, and the degree to which previous production trials primed subsequent verification trial errors and vice versa was measured. The main finding suggested that production errors were primed by previous production trials and verification errors were primed by previous verification trials, but neither type of error was primed by the previous trials with the other task. The findings also suggested that the problem difficulty effect was larger for the production task than the verification task. Taken together, these results imply that verification may reflect processes other than production; namely, it may involve Zbrodoff and Logan's resonance mechanism. One of the goals of the present study was to determine the extent to which verification is production plus comparison and the extent to which other strategies, such as resonance, are involved.

Multiple strategy use in verification

Although the idea of multiple strategies is not new to the discussion of verification task, there have been only limited efforts to formally implement multiple strategies into any of the proposed models of mental arithmetic because of a general lack of knowledge about the conditions in which each strategy is used. Ashcraft and Battaglia (1978) suggested

that retrieval could be "short circuited" by the presentation of a highly implausible answer, and Ashcraft and Stazyk (1981) theorized that the shorter RTs that they observed for large-split problems indicated that split information gave participants a way to bypass normal calculation. Another theory, first promoted by Krueger (1986) and more recently by Lemaire and Fayol (1995), states that in some cases participants can verify whether a given answer is correct by using the odd–even rule for multiplication. The odd–even rule states that if both operands of a simple multiplication problem are odd, then the product will be odd, and if one or both operands are even the product will be even. Because even–odd status of the operands determines the odd–even status of the product, participants should be, and were shown to be, faster and more accurate in rejecting differences between the given and correct answer of 1 or 3 than in rejecting differences of 2 or 4. Krueger (1986; cf. Ashcraft, 1982; Baroody, 1985; Campbell & Graham, 1985; LeFevre, Bisanz, et al., 1996) also reported that participants do not use the odd–even rule for equations with operands of 1, 0, or 5. This finding suggests that other, even simpler rules are available to bypass odd–even processing in some cases. Together these studies indicate that adults use a variety of strategies in verification.

Verbal protocols as data

Ericsson and Simon (1980, 1993) identified conditions under which verbal reports can yield valid data. Since then verbal reports have been used with general success throughout psychological research (Ericsson & Simon, 1980; for a review of many different areas, see Ericsson & Simon, 1993). It was not until recently, however, that investigators interested in simple numerical processing began to use verbal protocol methods. Two studies used verbal protocols to investigate processing in the production task, with some controversial results (LeFevre, Sadesky, & Bisanz, 1996; LeFevre, Bisanz, et al., 1996). For example, LeFevre, Sadesky, and Bisanz examined the problem difficulty effect in a simple addition production task. In this study, problem difficulty was more highly correlated with RTs on trials when participants reported a nonretrieval strategy than on trials when participants reported a retrieval strategy. On the basis of this finding, the authors suggested that the problem size effect and its implementation into models of simple arithmetic may be overemphasized. If this is true, it suggests that one cannot estimate the importance of problem size in any arithmetic task or implement it into a model without first determining the strategy used.

The main controversy in these studies pertains to the instructions given to participants. In the studies cited earlier the authors made references in the participant instructions to the idea that many different strategies

could be used to solve the problems. Furthermore, many of these strategies were explained to participants in detail. These instructions could have given participants insight into the purpose of the experiment and thereby increased the frequency of reporting and use of retrieval and nonretrieval strategies (or alerted them to strategies they might not otherwise have thought of). It is exactly for this reason that Ericsson and Simon (1980, 1993) discouraged the use of instructions detailing processes that might be involved in performing the task.

In another study, Kirk and Ashcraft (2001) presented data that suggest that the instructions used by the LeFevre group caused their participants to report and use nonretrieval strategies more often in the study using an addition production task (LeFevre, Sadesky, & Bisanz, 1996) and to report but not necessarily use this type of strategy more often in the study using production in multiplication (LeFevre, Bisanz, et al., 1996). Although these findings suggest that there may be a biasing effect of instructions in the LeFevre et al. studies, they are not a condemnation of the use of verbal protocols in general. In fact, the Ericsson and Simon (1980, 1993) framework for collecting verbal reports expects that this type of bias will result from suggestive instructions.

Much of the recent work addressing the use of multiple strategies in mental arithmetic have taken one of two tacks for avoiding the problems suggested by Kirk and Ashcraft (2001). In recent work Campbell and colleagues asked participants to classify their own processing into one of a few different categories (i.e., recognition, retrieve and compare, calculate and compare, odd–even rules, or other). This approach seems to be less susceptible to instructional biases (Campbell & Timm, 2000) and has been used to provide evidence that the greater number of odd operands and problem difficulty lead to decreased use of memory retrieval in simple addition production (Campbell, Parker, & Doetzel, 2004) and verification (Campbell & Fugelsang, 2001). This method has also been used to investigate RT differences in other mathematical operations and cultural groups (Campbell & Gunter, 2002; Campbell & Xue, 2001). Other researchers have continued to focus on performance measures to indicate the use of different strategies in different experimental situations. For example, the difference between RTs for problems that satisfy the odd–even rule and those that do not was found to vary with the number of odd operands included in a problem (smaller differences with fewer odd operands), as a function of the proportion of problems presented in a given experiment that violate the odd–even rule (larger differences when a majority of problems violate the rule), and as a function of practice (larger differences with more practice with a given problem set; Lemaire & Reder, 1999). Similarly, solution time differences were found between problems with 5 as operand (five-problems) and problems that do not

have 5 as an operand (non–five-problems). Furthermore, the difference between five-problems and non–five-problems was larger when the problem set included a larger proportion of five-problems (Lemaire & Reder, 1999).

In the present study we instructed participants to report their thoughts retrospectively, after doing the problem without any reference to possible strategies. We also instructed them to think of this as playing back a tape from the first thought they had at the presentation of the stimulus to the last thought they had before entering their response. The instructions are presented in Appendix A. As suggested by Ericsson and Simon (1980, 1993), these instructions should minimize the demand for participants to explain their actions, which has been shown to affect cognitive processing (Stinessen, 1985) and might have led to the bias in previous studies of mental arithmetic.

To summarize, we are interested in the use of verbal protocols to further the understanding of cognitive processes in mental arithmetic. Verification may be performed as a combination of production and comparison or through the use of completely different processes (i.e., a resonance or implicit retrieval process). The use of verbal protocols in verification should allow us to illuminate this relationship. Furthermore, the verification paradigm may allow us to gain insight into processes involved in simple arithmetic other than those available in production. Only by understanding the relationship between production and verification and determining whether the processing in two tasks is identical or overlapping can we develop a general theory of mental calculation.

In Experiment 1 we used a stimulus set that was used in Campbell's (1987b) primed production study to explore the distribution of strategies reported. It should be noted at the outset that problem difficulty and answer split were not orthogonally manipulated, such that, on average, hard problems also had small levels of split and easy problems had large split values. We used this stimulus set for the initial experiment because it was more important to first compare the behavioral effects from this study with those of the previous study by Campbell (1987b) to assess any change in cognitive processing that may result from collection of the verbal protocols.

EXPERIMENT 1

METHOD

Participants

Twelve students at the University of Colorado received course credit for their participation.

Apparatus and materials

Participants were seated at a table with a computer and tape recorder in front of them. Participants were asked to wear a headset microphone, through which their verbal responses during the whole experimental session were recorded. The experimenter was seated slightly behind and to one side of the participant.

Thirty-six single-digit multiplication problems were presented randomly four times in a blocked fashion across two experimental sessions. Across the four blocks, each problem was presented twice with an incorrect answer and twice with the correct answer. One of the incorrect answers was table related, and the other was table unrelated. All of the problem and answer combinations were used in an earlier study by Campbell (1987b).

Design

A 2 × 3 random block design was used. Both factors, problem difficulty and answer type, were inherent in the stimulus set (Campbell, 1987b). Problem difficulty was defined either as an easy or hard median split, based (as in Campbell, 1987b) on the normative RT data from Campbell and Graham (1985), or by using the same data as a continuous covariate. Answer type was defined by three levels: true, false and table unrelated, or false and table related.

Procedure

Participants were run individually in two 1-hr sessions separated by 3 days. In the first session, they initially ran through an alphabet verification task in order to get comfortable with the way the sessions were to be conducted and with reporting their thoughts. Participants were instructed that a pair of letters would appear at the center of the computer screen, and their task was to respond by pressing the key labeled "true" if the two letters were in alphabetical order. The letters did not have to be adjacent to each other in the alphabet; it was only necessary that the left-to-right ordering followed the before–after ordering in the alphabet. If the letters were determined to violate the before–after ordering of the alphabet, participants were instructed to respond by pressing the key labeled "false." Speed and accuracy were stressed equally in the instructions. For half the participants the "true" key was on the left, and for the other half the "true" key was on the right.

After the participants responded, they were prompted by a message on the screen to report the thoughts they remembered having while working on the problem from the first moment they saw the problem until they pressed the "true" or "false" key. Participants were asked to report their thoughts as specifically as possible and in the order in which they actually occurred. After the participants had reported their thoughts, the experimenter asked for any necessary clarification. After it was clear that the participants understood the task, the computer program was started, and participants proceeded through 24 trials of the alphabet task.

After completing the alphabet task, participants were given the instructions for the multiplication task. This task was similar to the alphabet task with the exception that multiplication problems were presented with candidate answers, and the participants were instructed to decide whether the given answer was true or false as quickly and accurately as possible. In both sessions there were two blocks of 36

problems in the verification task. The second experimental session was conducted exactly like the first except for the omission of the alphabet verification task.

RESULTS AND DISCUSSION

The results are presented in three main sections. The first two report errors and RTs to show that the results of the present study are consistent with published findings for verification. The third section is dedicated to the protocol analysis. Log base 10 transformations of the RTs were analyzed to reduce any outlier effects. All RT means are reported as anti-logs of the mean log values that were used for analysis (in all figures the log coordinates are maintained, but the numerical values are converted to anti-logs). In all analyses using the answer type variable, planned comparisons were performed between true problems and the average of all types of false problems and between table-related and table-unrelated problems. In all analyses using the two-level problem difficulty measure, a planned comparison between easy and hard problems was performed. These comparisons are of interest in comparing the present results with those of earlier studies (e.g., Campbell, 1987b; Zbrodoff & Logan, 1990).

Error data

A 3 (true, false and table unrelated, and false and table related) $\times$ 2 (two levels of problem difficulty) repeated-measures analysis of variance was performed on the proportion of incorrect responses in all trials. Participants made more errors on hard than on easy problems (7% vs. 3%), $F(1, 11) = 19.83$, $MSE = 0.0014$, $p < .01$. Planned comparisons for this analysis also showed that participants made fewer errors on problems that were true than the average of both types of false problems (3% vs. 6%), $F(1, 11) = 10.57$, $MSE = 0.0018$, $p < .01$. Furthermore, participants made more errors on problems for which the given answer was table related than on problems with a table-unrelated false answer (10% vs. 2%), $F(1, 11) = 18.38$, $MSE = 0.008$, $p < .01$. Finally, a significant interaction was found between problem difficulty and table-related and table-unrelated answers, such that the difference between the proportion of errors for the table-related and table-unrelated conditions was greater for hard problems than for easy problems, $F(1, 11) = 15.14$, $MSE = 0.0014$, $p < .01$. Specifically, participants averaged 1% and 6% errors for easy problems with table-unrelated and table-related answers, respectively, but averaged 3% and 13% errors for hard problems with table-unrelated and table-related answers, respectively.

RT data

The analysis performed for log RTs included only the trials for which participants' responses were correct and used the same design as the er-

ror analysis reported earlier. Averaged across answer type, participants were slower to respond to harder problems (M = 1,638 ms) than on easy problems (M = 1,273 ms), $F(1, 11) = 42.50$, $MSE = 0.005$, $p < .01$. Averaged across problem difficulty, participants were faster responding to true problems (M = 1,299 ms) than to the average of both types of false problems (M = 1,523 ms), $F(1, 11) = 40.83$, $MSE = 0.0037$, $p < .01$. Furthermore, averaged across difficulty, participants were faster responding to false problems that were table unrelated (M = 1,455 ms) than to false problems that were table related (M = 1,593 ms), $F(1, 11) = 19.00$, $MSE = 0.0019$, $p < .01$, for log RTs.

The results of the error and log RT analyses are consistent with Campbell's (1987b) data obtained in a primed production task and with other published studies that have manipulated problem difficulty in the verification paradigm (Zbrodoff & Logan 1990) and relatedness in verification (Koshmider & Ashcraft, 1991; Stazyk, Ashcraft, & Hamann, 1982). These replications suggest that protocols did not affect performance in this task in any significant way.

Protocol analyses

Verbal protocols were coded separately by two different coders. Disagreements between the coders were then resolved through argument, and if no agreement could be reached the trial was subsequently coded as uninterpretable and not included in further analyses. Thus, the trials that were used in the analysis of the protocols were ones in which there was 100% agreement between the coders. Each coder placed each trial into one of 17 report categories. Some of the categories corresponded to a priori theoretical hypotheses based on the literature. These categories included retrieve–compare, calculate–compare, pattern match, magnitude estimation, and other rules that have been suggested in the mathematical cognition literature (e.g., five-problem and odd–even rules). Other categories were created to group similar protocols together that did not fit into any of the a priori categories (e.g., uninterpretable and recency effects) or to code for other characteristics of the protocols (e.g., explicit no answer generation or operand switch). Trials in which participants made references to multiple strategies were coded with multiple codes, with the first being the strategy that the coder considered the one most likely to lead to the participant's response (i.e., the dominant mode of processing for that trial). Because there were few trials with multiple codes, and we were more interested in differentiating between trials that used only one particular strategy, multiply coded trials were not further analyzed. Appendix B describes all 17 categories.

Evidence of multiple strategies in verification. Table 1 presents the proportion of the trials on which each of the most frequent strategy cat-

Table 1. Raw frequency, proportion of total trials reported, and mean response time for each report category

Report category	Raw frequency, Experiment 1	Raw frequency, Experiment 2	Proportion of trials, Experiment 1	Proportion of trials, Experiment 2	Mean RT (ms), Experiment 1	Mean RT (ms), Experiment 2
Retrieve–compare	1,154	1,682	.67	.55	1,442	1,951
Calculate–compare	117	96	.07	.03	2,269	2,256
Pattern match	135	529	.08	.17	1,443	1,568
Magnitude estimation	90	234	.05	.08	1,464	1,639
Reverse retrieve–compare	60	182	.03	.06	1,871	1,494
Odd–even rule	4	9	.002	.003	1,006	1,347
Other	168	321	.1	.1		

egories was reported and the corresponding mean RT. It is important to note that all participants reported a mixture of strategies.

The proportion of trials on which participants reported the retrieve–compare strategy was analyzed in a 2 (problem difficulty: easy, hard) × 3 (answer type: true, false related, false unrelated) ANOVA. Strategy reports were categorized as "retrieve–compare" when participants stated that they had retrieved the correct answer from memory and compared it with the presented answer. In replication of previous studies (Campbell & Xue, 2001; LeFevre, Sadesky, & Bisanz, 1996), participants reported using the retrieve–compare strategy more on easy problems than on hard problems (72% vs. 63%), $F(1, 11) = 4.83$, $MSE = 0.0283$, $p = .05$. Note that the retrieve–compare category is synonymous with production plus comparison. Thus, this first analysis of the verbal protocols suggests that production plus comparison is the modal strategy, in contrast to the findings of Zbrodoff and Logan (1990) and Campbell and Tarling (1996). Verification, however, is not performed purely through the use of this strategy.

Validation of strategy categories.

To confirm that the four most frequently reported categories represent different strategic processing, the following analyses focused on finding specific differences in RTs or some sort of regularity regarding the problems or conditions to which the strategies were applied. It is important to note that although different processing as defined by the strategy categories might not necessarily mean that the processes themselves take different amounts of time, the RT differences we sought were motivated by previous research and theory regarding the hypothesized processes (Ashcraft & Stazyk, 1981; Campbell & Tarling, 1996; Zbrodoff & Logan, 1990).

In the following analyses, Campbell and Graham's (1985) normative data (i.e., adults' mean correct RT from this study) were used as a continuous measure of problem difficulty to allow the assessment RT differences resulting from strategy and answer type while still controlling for problem difficulty. Coding problem difficulty as a continuous covariate produced a reduction of the design matrix while still allowing statistical control of problem difficulty in the analyses. This type of statistical control of problem difficulty (as opposed to the two-level variable used in previous analyses) was necessary because only participants with trials in each cell of the design could be used in these analyses. Therefore, it is important to control for problem difficulty in these analyses because we have no control over which strategies are used for which problems. One consequence of this approach, however, is that main effects and interactions with the continuous problem difficulty measure might be difficult to interpret. Finally, although these analyses were limited to participants who reported each

strategy in all cells of the design, it is important to note that the number of participants used was not the number of participants who reported a particular strategy but of those who reported using the strategy for each type of problem (i.e., true, false related, and false unrelated). Indeed, all the categories we analyzed were reported at least once by the majority of the participants. Finally, with regard to the problem difficulty measure, it is important to point out that Campbell and Graham's study used a pure production task, and at least two studies have suggested that the problem difficulty effect is greater for production than for verification (Campbell, 1987b; Campbell & Tarling, 1996). The larger problem difficulty effect for production does not, however, invalidate the use of the problem difficulty measure in the present study because Campbell (1987b) found the correlation between production and true verification RTs to be quite high.

A 2 (strategy: calculate–compare or retrieve–compare) × 3 (answer type: true, false related, or false unrelated) repeated-measures analysis of covariance was performed for trials in which participants' responses were categorized as either retrieve–compare or calculate–compare. Trials were categorized as calculate–compare when participants stated that they had used an intermediate calculating algorithm to produce the correct answer, which they compared with the answer given. Because using a calculating algorithm implies more processing than simple retrieval, RTs for the calculate trials should be longer than those for the retrieve–compare trials, and this difference should be more pronounced for difficult problems. Four participants were used in the analysis. The data are presented in Figure 1.

RTs were shorter for true problems than for the average of both types of false problems, with problem difficulty controlled, $F(1, 3) = 21.73$, $MSE = 0.00144$, $p = .04$, and participants also responded more slowly to problems when they reported using calculate–compare, with problem difficulty controlled, $F(1, 3) = 473.02$, $MSE = 0.00004$, $p < .01$.[1] One possible reason for this finding is that participants use the calculate–compare strategy when they fail to retrieve the correct answer while trying to apply the retrieve–compare strategy. In general, however, calculation can be expected to be slower than retrieval because calculation usually implies more steps to reach a conclusion than does one-step retrieval (Baroody, 1985; Rickard, 1997). For example, a rule that was commonly reported for problems with 9 as one of the operands was to retrieve the answer to 10 times the other operand and then subtract that operand from the result. This rule entails two steps to ascertain the correct answer before comparison to the answer given, which should generally take longer than just one retrieval step.

Trials categorized as magnitude estimation are those for which participants reported that the answer was either too large or too small to be

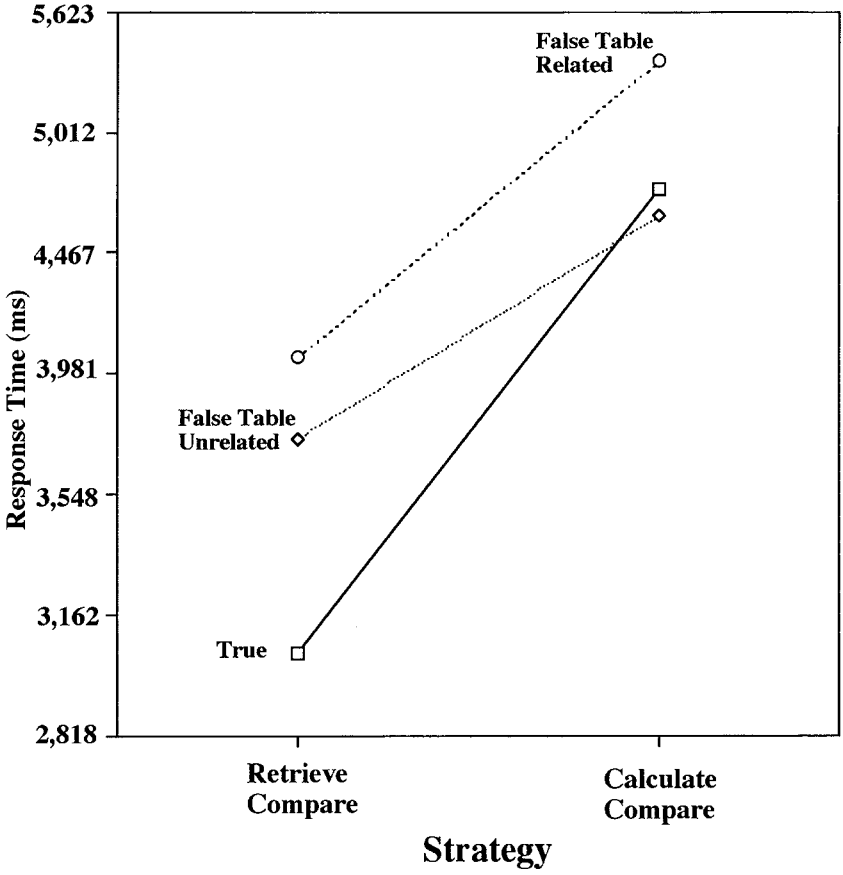

Figure 1. Anti-log mean response times for retrieve–compare and calculate–compare strategies by answer type. Means are adjusted for the problem difficulty covariate

correct. On many of these trials participants spontaneously reported that they did not know what the correct answer was, but they knew that the given answer was either too large or small. The first analysis of these data looked for log RT differences between the trials categorized as magnitude estimation and those categorized as retrieve–compare while controlling for answer type, problem difficulty, and the difference between the correct answer and the answer given. Welford's similarity function, defined in Campbell and Oliphant (1992),[2] was used as the measure of difference between the given answer and the correct answer (i.e., split). Because the Welford value (i.e., log[larger/(larger − smaller)]) would be undefined for all true problems, and the use of the magnitude estimation strategy

for true problems is highly improbable, correct problems were omitted from this analysis. Only the data for the five participants who used the magnitude estimation strategy for five or more trials were used for these analyses.

If the magnitude estimation strategy allows participants to bypass normal (retrieve–compare) processing, then the RTs should be shorter for the trials categorized as magnitude trials. In contrast, if the magnitude estimation strategy is used in cases where the correct answer is attempted but fails, the RTs should be longer for magnitude estimation trials. The results of this analysis yielded no significant differences in log RTs.

Given the outcome of the analysis comparing RTs for the retrieve–compare and magnitude estimation strategy categories, a second analysis was performed on the magnitude estimation trials. This analysis focused on changes in the proportion of use of the magnitude estimation strategy as a function of split. For this analysis trials were grouped into five equally spaced levels of split, with "1" indicating *small split* and "5" indicating *large split*. Linear through quartic components were evaluated on the proportion of trials on which participants reported using the magnitude estimation strategy. True trials were omitted from this analysis for the same reason as in the RT analysis of magnitude estimation trials, and the same five participants' data used in that analysis were included here. Regardless of whether the magnitude estimation strategy is used to sidestep answer retrieval or as a backup when retrieval fails, the ease or frequency of use of the magnitude estimation strategy should increase with split. In fact, participants' reported use of the magnitude estimation strategy increased linearly with increasing levels of split (i.e.,. Welford values), $F(1, 8) = 9.48$, $MSE = 0.20303$, $p = .02$. This effect is shown in Figure 2.

The final analyses compared the retrieve–compare and pattern match strategies, with problem difficulty as the covariate. For trials coded as pattern match, participants stated that the problem just looked right or wrong, with no intermediate steps or calculations. We believe that this strategy may be synonymous with resonance processing (as described by Zbrodoff & Logan, 1990) because participants reported a matching-like procedure, often stating that they did not retrieve the correct answer but responded based on how the problem looked. This definition is similar to the recognition category used by Campbell and colleagues (e.g., Campbell & Fugelsang, 2001) and is consistent with the idea that production and verification use different memory processes (i.e., Campbell & Tarling, 1996). Because of the explicit matching procedure we expected that these were not just trials in which they had lost access to intermediate processes from short-term memory. We also expected that when participants lost access to the results in short-term memory, they would report that they either did not know how they arrived at the answer or would report that

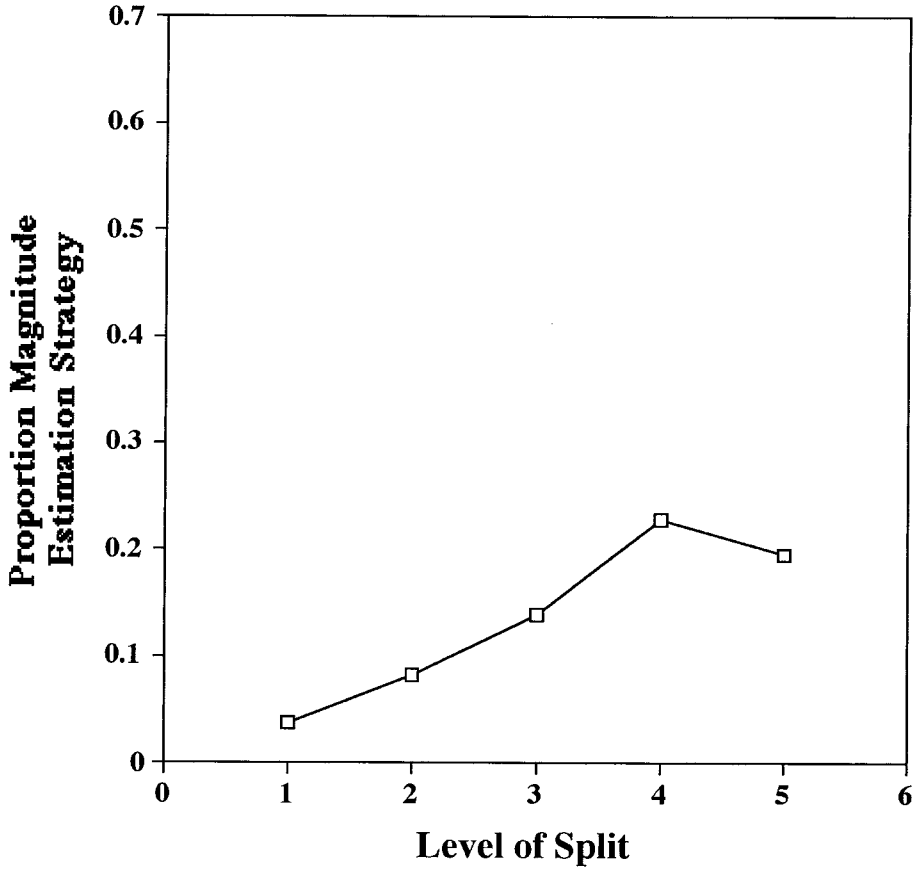

Figure 2. Mean proportion of trials using magnitude estimation strategy by level of split

they forgot. This analysis, however, yielded no significant results, which is inconsistent with Zbrodoff and Logan's (1990) suggestion that participants compare the whole equation with an earlier instance of the problem.

EXPERIMENT 2

In Experiment 2 we sought to clarify several questions concerning the authenticity of calculate–compare, magnitude estimation, retrieve–compare, and pattern match strategies and to further investigate the conditions in which they are applied. First, we sought to replicate findings from Experiment 1 concerning the use of the retrieve–compare and calculate–compare strategies. Second, we wanted to investigate RT differences

between retrieve–compare and magnitude estimation trials when we orthogonally manipulated problem difficulty and split in the stimulus set. Also concerning the authentication of the magnitude estimation strategy, we wanted to determine whether the pattern of increasing use of the magnitude estimation strategy with larger split would replicate when difficulty and split were not confounded. Finally, we expected that Experiment 2 would give us additional information about the pattern match strategy.

We made several adjustments in Experiment 2 to increase the number of overall observations and to manipulate split and problem difficulty in an orthogonal manner. Each problem was presented eight times: four times with a correct answer, once each with table-related false answers of large and small split, and once each with table-unrelated answers of large and small split. In addition, we increased the number of participants in Experiment 2.

METHOD

Participants

Sixteen introductory psychology students at the University of Colorado received course credit for their participation.

Apparatus and materials

Participants were seated at a table with a computer and tape recorder in front of them. Participants were asked to wear a headset microphone, and the experimenter was seated to one side.

The problem and answer set was based on that used in Experiment 1, with the exception that some easy problems with operands less than 3 and small squares up to and including 5×5 were omitted. Two additional types of answer primes were also included to balance the difficulty and split features. Specifically, 24 single-digit multiplication problems were presented eight times over three sessions. Four presentations contained the true answer and four contained a false answer. One false answer was unrelated to the multiplication table of either operand and of small split. A second was unrelated and of large split, and a third and fourth were related and of small and large split, respectively (as indicated by Welford's similarity function). The mean Welford values and their standard deviations are given in Table 2 for each false answer problem.

Table 2. Mean Welford (i.e., similarity) values and their standard deviations for each type of problem

	Table unrelated		Table related	
	M	*SD*	*M*	*SD*
Small split (i.e., high similarity)	1.32	0.296	0.89	0.088
Large split (i.e., low similarity)	0.32	0.041	0.46	0.076

Design

A 3 × 2 × 2 random block design was used for this experiment. The first factor, answer type, was defined by the properties of the answer prime presented in each trial (true, table related, or table unrelated). The second factor, split, was defined within the two false answer types such that one of the two presentations of each type of false problem included an answer that was a small numerical distance from the correct answer (i.e., small split) and one presentation was with an answer that was a large numerical distance from the correct answer (i.e., large split). The third factor, problem difficulty, consisted of two levels based (as in Experiment 1) on median split of normative RT data from Campbell and Graham (1985). To illustrate this classification, two problems and the five presented answers are listed in Table 3. The first problem is an example of a hard problem, and the second problem is an example of an easier problem. For the complete problem set, see Appendix C. All false answers were nonprime, viable answers to other simple multiplication problems. Note that because of this constraint and the need to keep split consistent across all problems, the odd–even rule could be applied to verify all false problems. Even so, we expected few applications of the odd–even rule, given its infrequent reported occurrence in Experiment 1.

Procedure

The procedure for Experiment 2 was the same as that for Experiment 1 with one exception. Participants in Experiment 2 completed three sessions, and the third session was conducted exactly like the second.

RESULTS AND DISCUSSION

Error data

Overall accuracy was high, and on average the participants never made more than 10% errors in any condition. Thus, any analysis of the errors is not reported because of the possibility of contamination by ceiling effects. Errors are reported for completeness in Table 4.

RT data

The first analysis was a 3 (answer type: true answers, false table-unrelated answers, and false table-related answers) × 2 (split: high values, indicating large differences, and low values, indicating small differences between the

Table 3. Example problems

Problem	True	False, unrelated, small split	False, unrelated, large split	False related, small split	False, related, large split
7 × 9	63	64	32	56	42
3 × 9	27	28	14	24	36

Table 4. Anti-log response time means (ms), proportion of errors (PE), and their respective standard deviations for Experiment 2

	Easy				Hard			
Problem type	RT	*SD* (log)	PE	*SD*	RT	*SD* (log)	PE	*SD*
True	1,051.96	.10	.02	.03	1,336.37	.15	.07	.05
Unrelated, small split	1,289.42	.13	.07	.13	1,543.20	.18	.07	.06
Unrelated, large split	1,284.56	.13	.03	.08	1,408.07	.17	.05	.07
Related, small split	1,355.20	.13	.06	.09	1,625.92	.19	.10	.06
Related, large split	1,370.35	.15	.02	.05	1,499.43	.18	.04	.06

given and correct false answers) × 2 (problem difficulty: easy and hard) repeated-measures ANOVA with log RTs as the dependent variable. The anti-log means and standard deviations for each cell in the design are presented in Table 4.

As has been found repeatedly (Experiment 1; Campbell, 1987b; Koshmider & Ashcraft, 1991; Stazyk et al., 1982; Zbrodoff & Logan, 1990), participants in Experiment 2 were slower to verify false problems than true problems, $F(1, 15) = 31.16$, $MSE = 0.0049$, $p < .01$, and slower to verify problems that were presented with table-related false answers than table-unrelated false answers, $F(1, 15) = 25.00$, $MSE = 0.0007$, $p < .01$. Participants also were slower to verify hard problems than easy problems, $F(1, 15) = 21.90$, $MSE = 0.0084$, $p < .01$, and in replication of Campbell and Tarling's (1996) findings, the RT difference between true and the average of all types of false problems was smaller for hard problems than for easy problems, $F(1, 15) = 13.99$, $MSE = 0.0009$, $p < .01$. As in Experiment 1, the replication of these effects suggests that the introduction of retrospective protocols to the verification task did not alter it in any substantive way.

The new and more interesting outcomes of Experiment 2 are that participants took longer to verify problems that were presented with small split answers than those presented with large split answers, $F(1, 15) = 8.02$, $MSE = 0.001$, $p = .02$, and a significant interaction of split and problem difficulty was found such that, on average, the RT difference between problems presented with small and large split answers was larger for hard problems than easy problems, $F(1, 15) = 8.81$, $MSE = 0.001$, $p < .01$. The influences of split on the RTs and errors support Ashcraft and Stazyk's (1981) findings and demonstrate the need to control for these effects to get the purest picture of the verification task and the factors involved in it. Zbrodoff and Logan (1990) used the findings of Ashcraft and Stazyk as evidence that production plus comparison is not the sole processing in verification. Zbrodoff and Logan suggested that split effects indicate that participants may evaluate the equation as a whole and make their decision without computing or retrieving an answer. That is, participants

may determine whether the answer is plausible for the given problem. The split effects in the present data support the interpretation that production plus comparison is not the unitary verification process. Although, as discussed later, they do not necessarily imply the use of resonance.

Protocol analyses

Each trial in this experiment was categorized by two coders into one of the 17 different report types identified in Experiment 1, and the same procedures were used to handle disagreements between the coders and multiple strategies. Thus, these report categories did not need further development for use in Experiment 2 even though they were derived from the independent data of Experiment 1. This fact supports the veridical nature of the report categories.

Evidence for multiple strategies in verification. The 17 categories are listed in Appendix B. The frequencies of occurrence of the most frequent categories, along with the mean RTs and proportion of trials for which each strategy was reported for are listed in Table 1. In replication of Experiment 1, the majority of participants reported a mix of strategies.

The first of the protocol analyses investigated trials on which participants' verbal reports were categorized as retrieve–compare. This analysis was identical to that used in Experiment 1. Participants reported using the retrieve–compare strategy less often for hard problems than for easy problems (67% vs. 53%), $F(1, 15) = 18.11$, $MSE = 0.0466$, $p < .01$. Unlike in Experiment 1, the difference in the proportion of use of the retrieve–compare strategy between true and false problems was greater for easy problems than for hard problems, $F(1, 15) = 8.87$, $MSE = 0.0116$, $p < .01$. This effect is shown in Figure 3. More interesting for current purposes, participants reported using the retrieve–compare strategy less often when the problems were presented with large-split answers than with small-split answers (55% vs. 67%), $F(1, 15) = 14.16$, $MSE = 0.0294$, $p < .01$.

The results of this analysis provide additional insight into the influences of split in verification. In replication of Experiment 1 and previous studies (Campbell & Xue, 2001; LeFevre, Sadesky, & Bisanz, 1996), participants reported using the retrieve–compare strategy less often for hard problems than for easy problems, but in this experiment participants also reported this strategy less often for problems that were presented with large-split answers. In conjunction with the RT results, the influence of split on the proportion of retrieve–compare trials also supports the notion that verification involves more than production plus comparison.

The effects of problem difficulty and split on the proportion of trials in which participants used the retrieve–compare strategy imply that strategy choice in this task is influenced in part by problem structure. Specifically, if the difference between a given false answer and the correct answer is

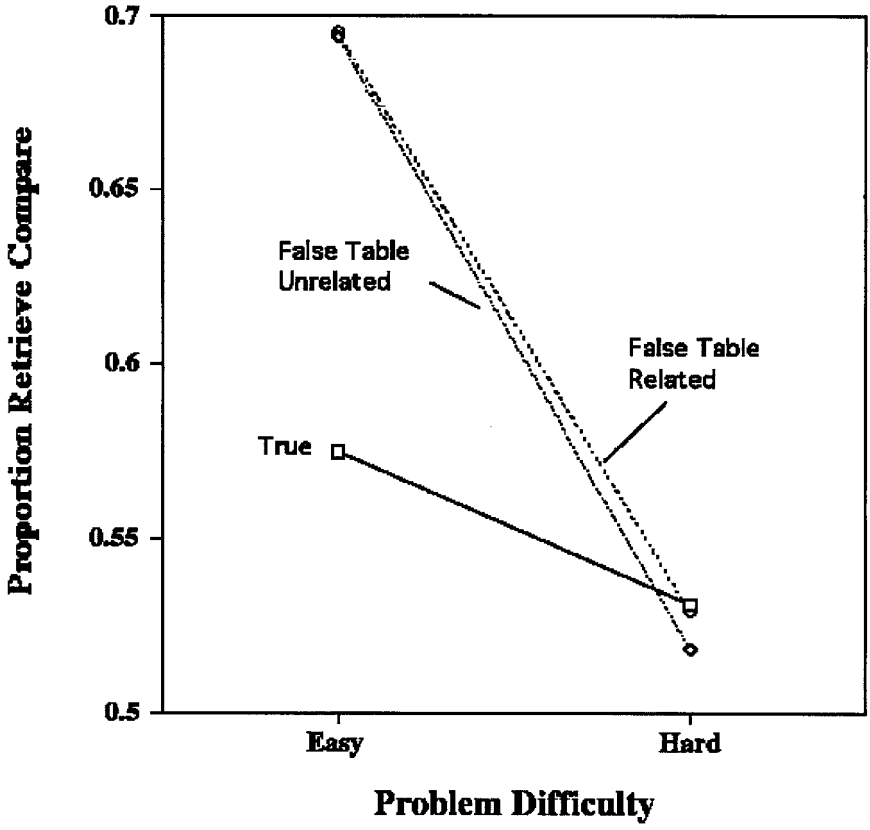

Figure 3. Mean proportion of trials categorized as retrieve–compare for true and false problems by level of problem difficulty

large, then participants are more likely to choose a strategy that uses split information for verification (e.g., the magnitude estimation strategy). If, however, the difference between the given and the correct answers is small, participants are less likely to rely on split and more likely to choose retrieve–compare, calculate–compare, or some other strategy that is not based on the split information (e.g., a $\times 5$ rule or a $\times 9$ rule, if possible). Although participants may choose production-like strategies such as retrieve–compare or calculate–compare for problems that are presented with small-split answers, they are less likely to use them for difficult problems. Thus, when hard problems are presented with small-split answers, the use of retrieval is also less likely than the use of either calculate–compare or a sidestepping strategy such as the $\times 5$ or $\times 9$ rules.

Validation of strategy categories. The proportion of trials in which participants reported using the magnitude estimation strategy was analyzed as

a function of split (measured by Welford's similarity function). Split values were grouped into five equally spaced levels, with "1" indicating *small split* and "5" indicating *large split*. Linear through quartic trend components were evaluated on the proportion of trials in which participants reported using these strategies. The data, presented in Figure 4, were derived from 12 participants who reported using the magnitude estimation strategy on five or more trials. In replication of Experiment 1, the prediction that participants would use the magnitude estimation strategy more often as split got larger was confirmed. Both the linear and quadratic trend components were present such that as split became larger, participants reported using the magnitude estimation strategy more often, $F(1, 11) = 38.86$, $MSE = 0.0124$, $p < .01$ for the linear trend and $F(1, 11) = 15.82$, $MSE = 0.0106$, $p < .01$ for the quadratic trend.

A comparison of magnitude estimation trials with retrieve–compare trials in log RTs was the focus of the next analysis. Campbell and Graham's (1985) continuous measure of problem difficulty was used as the covariate. The other factors in the analysis were two levels of answer type (false, table related and false, table unrelated), two levels of split and two strategy categories. The data from four participants who had observations in each

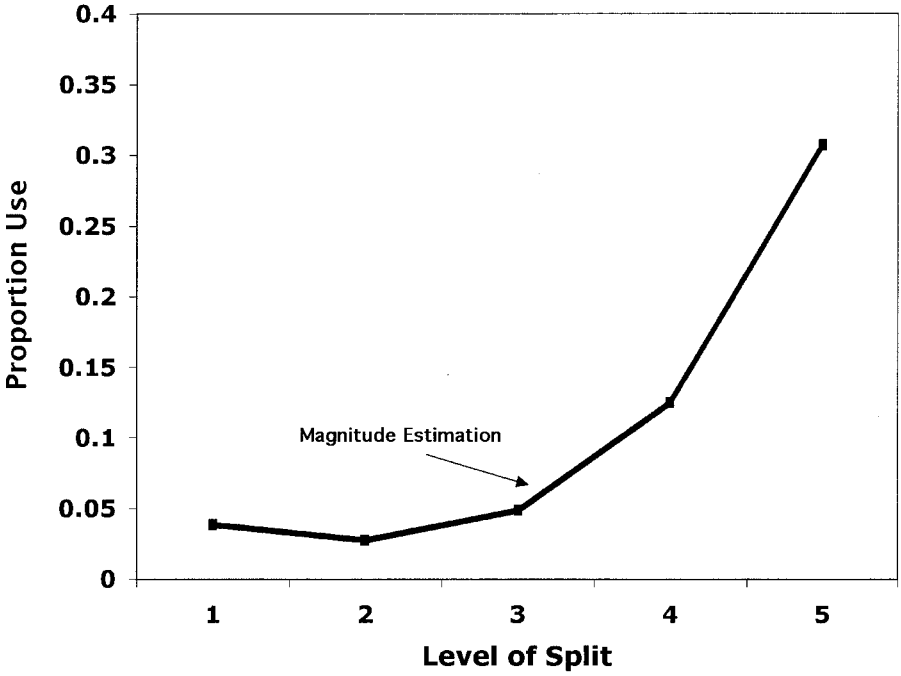

Figure 4. Mean proportion of trials categorized as magnitude strategy by level of split

cell of this design were used in the analysis. The magnitude estimation strategy could be used in either of two contrasting situations: to bypass normal (retrieve–compare) processing, which would be indicated by shorter RTs for the magnitude estimation trials than for retrieve–compare trials; and after retrieval fails, which would be indicated by longer RTs for the magnitude estimation trials. The analysis of covariance showed that trials categorized as magnitude estimation were faster than retrieve–compare trials, with problem difficulty controlled, $F(1, 3) = 27.42$, $MSE = 0.0007$, $p = .03$. The RT difference between magnitude and retrieve–compare trials was also greater for large-split problems than for small-split problems, controlling for problem difficulty, $F(1, 3) = 24.34$, $MSE = 0.0004$, $p = .04$. These effects are shown in Figure 5.[3]

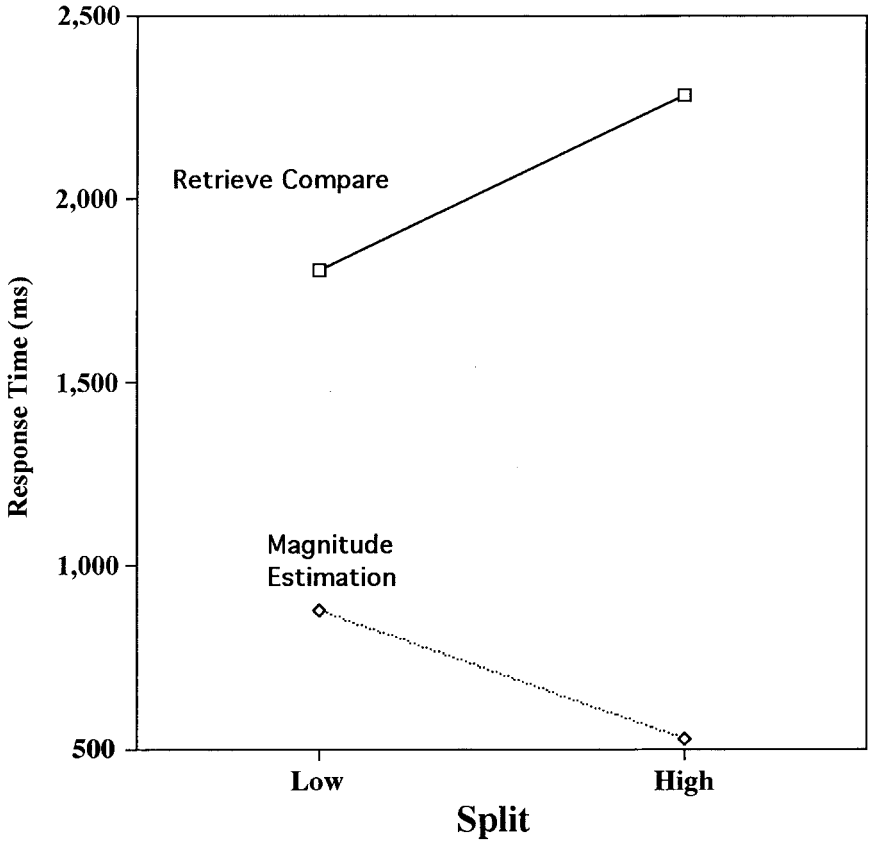

Figure 5. Anti-log mean response times for trials categorized as retrieve–compare and magnitude estimation by level of split. Means are adjusted for the problem difficulty covariate

It follows from these findings that participants do make plausibility judgments based on the answer and operands and that these plausibility judgments are used as a way to bypass normal retrieval or calculation of the correct answers. This interpretation is also consistent with the fact that when reporting magnitude estimation, participants often did not think about the correct answer. Furthermore, participants' ability to bypass normal retrieval or calculation processes suggests that retrieval may not occur automatically, regardless of the participants' intentions. Again, the existence of sidestepping strategies supports the notion that verification involves both production plus comparison and sidestepping operations.

A comparison of retrieve–compare trials with pattern match trials was the focus of the next analyses. A 3 (answer type) $\times$ 2 (split) $\times$ 2 (strategy) ANCOVA was conducted with the continuous measure of problem difficulty as the covariate. Six participants contributed data for this analysis. If the pattern match strategy involves no calculation or retrieval, the RTs for those trials should be shorter than for retrieve–compare trials, and the difference should be more pronounced for true problems. Although participants were faster overall to verify true than false problems (true = 1,188 ms, false = 1,282 ms), $F(1, 5) = 59.03$, $MSE = 0.0006$, $p < .01$, and there was an interaction between the true–false contrast and the parallel coded difficulty contrast, $F(1, 5) = 19.59$, $MSE = 0.0006$, $p = .01$, there were no differences in RTs between the two strategy categories and no interactions that included the strategy factor. Thus, the hypothesis that pattern match and retrieve–compare trials represent quantitatively different strategies or retrieval mechanisms did not find much support in the data of Experiment 2. Indeed, the significant effects from this final analysis are completely redundant with those of the overall RT analysis reported earlier. The consistency of the findings of this analysis with the overall RT analysis suggests that the trials categorized as pattern match are simply fast retrieve–compare trials.

GENERAL DISCUSSION

The present study has provided evidence through the use of retrospective protocols that should help to clarify several issues about tasks commonly used in investigations of mental calculation. Our findings support participants' use of multiple strategies in mental multiplication. In the present experiments we used instructions that made no explicit references to how the task could be performed. In contrast to the LeFevre, Bisanz, et al. (1996) study and consistent with the Ericsson and Simon (1980, 1993) framework for collecting verbal reports as data, we instructed participants to report the thoughts that they remembered having after performing each trial of the task. According to studies reviewed by Erics-

son and Simon, soliciting verbal reports in this manner should preclude any demand effects that are similar to those reported by Kirk and Ashcraft (2001). As pointed out earlier, the Kirk and Ashcraft findings are not a complete condemnation of the use of verbal protocols in the study of simple multiplication but rather a cautionary tale regarding how not to collect this type of data.

By adhering to the Ericsson and Simon (1980, 1993) framework for collecting verbal reports, we predicted that the cognitive processes involved in these experiments would be unaffected. Our replications of the patterns of effects reported previously in the mathematical cognition literature support the assertion that fundamental cognitive processes involved in mental calculation were unchanged by the requirement that participants provide verbal retrospective protocols. The agreement of the behavioral measures with reported strategies also supports the validity of these reports and of the categories they yield.

Verification and production

Our evidence suggests that verification is not performed solely through production plus comparison but rather that verification is performed through a mixture of production plus comparison and sidestepping strategies.

In contrast to the findings of Campbell and Tarling (1996), the dominance of the production plus comparison strategy suggests that verification and production tasks are not based on different processes. Indeed, the present data suggest that depending on the makeup of the stimulus set, verification can be viewed as primarily production plus comparison in some cases. Although many studies have found a larger problem difficulty effect for retrieval-based strategies than for procedure-based strategies (Campbell et al., 2004; Hecht, 1999; LeFevre & Morris, 1999; Robinson, 2001), the larger problem difficulty effect for production found in the Campbell and Tarling (1996) study may have resulted from the mixture of strategies specifically possible in the verification task. The verification task enables strategies such as magnitude estimation and 5 or 0 rules that do not necessitate any memory retrieval of an answer. Problem difficulty might have a smaller effect on these types of sidestepping strategies than on retrieve–compare and calculate–compare processing, or even no effect. This would decrease the overall problem difficulty effect in verification, even with the primary use of production-like processing in the task. This interpretation is consistent with previous studies that suggest that the problem difficulty effect reflects disruption in the memory retrieval stage and not in the encoding or response stages of performance (Campbell & Clark, 1992; Campbell & Fugelsang, 2001).

What are the conditions under which production and verification

rely on the same processes? The change in the distribution of strategies between the two experiments suggests that the proportion of trials for which production plus comparison accounts in verification should depend strongly on the features of the problem set. Production plus comparison (i.e., retrieve–compare) accounted for a larger proportion of the trials in Experiment 1 than in Experiment 2. Similarly, participants' reported use of the pattern match and magnitude estimation strategies increased in Experiment 2, and the reported use of the calculate–compare strategy declined. Clearly these differences must result from the changes in the problem set because all other factors remained the same in Experiments 1 and 2. One way to interpret these qualitative changes is to suggest that verification can be set up so that the processing is very similar to production. If we assume for the moment that production is based primarily on memory retrieval, these data suggest that verification can be controlled to reflect a similar basis. The present data suggest that this can be accomplished by using easy problems and false answers with very small splits that are related to the correct answers. This interpretation is consistent with previous studies that have found increases in use of a particular strategy with problem sets that include a larger proportion of a particular type of problem (Lemaire & Reder, 1999)

Strategies in verification

By far the strongest support was for the use of the magnitude estimation strategy for certain types of problems. This strategy was reported more often as split became larger, and in Experiment 2 magnitude estimation trials were faster than retrieve–compare trials. These findings support the sidestepping nature of this processing that capitalizes on the availability of split information. It may be useful to point out that although there were no RT differences between the retrieve–compare and magnitude estimation trials in Experiment 1, participants reported the magnitude estimation strategy more often as split increased. We can speculate that magnitude processing may even be used for smaller splits but that RT measures are not sensitive enough to detect it in these cases. The redundancy between split and problem difficulty in this stimulus set may have masked any RT differences between the strategies because magnitude estimation probably would be used only when split is large, which in the case of Experiment 1 was on easy problems for which retrieval times already were short. In this case there might not be a difference between retrieve–compare and magnitude estimation trials because the retrieve–compare trials are at the fast end of their respective distribution, which may be similar to the times necessary to use split information in the magnitude estimation trials. This possibility further underscores the value of retrospective reports to study mental arithmetic.

Another strategy supported by the data is the calculate–compare strategy. In Experiment 1, participants were slower to verify problems when they reported the use of a calculation algorithm. This analysis, however, was not possible in Experiment 2 for lack of power. The calculate–compare strategy was reported for a smaller proportion of the trials in Experiment 2 than in Experiment 1, which suggests that split and difficulty also are involved with the selection of this strategy. Specifically, the use of calculate–compare is facilitated when these factors are manipulated in a manner similar to that in Experiment 1, but when split and problem difficulty are balanced, as in Experiment 2, other strategies take precedence. A specific hypothesis based on this speculation would predict facilitation of the calculate–compare strategy as conditions were constructed to allow verification to become more similar to production.

In the present study, we believe that the pattern match report category best resembles the use of resonance or degree of match as described by Zbrodoff and Logan (1990). There are several reasons to expect RT differences between retrieve–compare and pattern match trials, and so the lack of RT difference between these strategies is inconsistent with what should be quantitatively different processing. Furthermore, if the overall RT analysis is driven by the primary use of retrieve–compare processing and there is no difference between the underlying processing involved in the retrieve–compare and pattern match strategies, we would expect the similar effects for the analysis including these two strategies and the overall RT analysis. Based on the lack of differences between these trials and the similarity between the analyses noted earlier, we speculate that pattern match trials are instances of retrieve–compare trials in which processing is too fast for the deposit of any verbalizable results in short-term memory, or the deposits in short-term memory fade too quickly for later report. For whatever reasons this occurs, the absence of information in short-term memory results in a lack of cues from which participants could reconstruct the processes involved at the time of the verbal report. For these reasons we are not optimistic about theories of verification that revolve around the use of resonance, but let us consider several other points of view.

Originally, Zbrodoff and Logan (1990) suggested that split effects in verification indicate that participants compare the equation as a whole without retrieving or computing an answer and that the degree of match (resonance) was the criterion for a verification decision. Thus, it could be argued that our definition of the pattern match trials does not characterize the type of processing that Zbrodoff and Logan (1990) had in mind. Furthermore, it could be argued that the processing we have labeled as magnitude estimation better characterizes their ideas. We can think of one reason why this argument may not be the case. Based on their

original description of resonance (Zbrodoff & Logan, 1990), we believed that use of this strategy should be accompanied by information in short-term memory (for later verbal report) about matching the equation as a whole without any reference to specific features. In contrast, we expected that the use of the magnitude estimation strategy would deposit some information (in short-term memory) concerning the size of the answer in relation to the operands and that participants would report this information in their protocols. This prediction was well supported in all of the magnitude estimation protocols because participants always reported that the answer was either too small or too large for the problem. It may well be that resonance processing is synonymous with what we have labeled magnitude estimation, but further study or formal implementation of resonance into a working model may be necessary to illuminate any possible similarities.

Dual-process memory theories assert that recognition responses can be based on recollection (i.e., retrieval of specific information) or familiarity (for review, see Richardson-Klavehn & Bjork, 1988; Roediger & McDermott, 1993). According to this work, recollection is characterized as a slow search process relying on associations, attention, and conceptual processing. In contrast, familiarity is believed to be a faster process (Atkinson & Juola, 1974; Jacoby, 1991; Mandler, 1980) that relies on a match of perceptual characteristics (Jacoby & Dallas, 1981) and does not require attention (Jacoby, 1991). Campbell and Tarling (1996) discussed resonance processing in mental arithmetic in terms of familiarity and suggested that verification is based primarily on familiarity of the candidate answer (equation), whereas production is based on retrieval of a candidate product. Although such an argument supports our unrealized expectation of RT differences in the present study, adoption of the methodology of recognition memory studies might better illuminate the use of familiarity in mental calculation. More specifically, it may be that our use of verbal protocols is not sensitive enough to differentiate familiarity from recollective processing. The use of the process dissociation procedure (Jacoby, 1991) or remember–know methods (Gardiner, 1988) may allow better distinction and understanding of resonance processing in mental arithmetic. We suggest that the use of these procedures may be useful in future studies of mental arithmetic.

Explicit reports of other candidate sidestepping strategies could not be analyzed because of low frequencies of use (an inspection of Appendix A will yield a general description of other candidate sidestepping strategies). Some of these strategies have been investigated in the literature, and some appear to be unique to our data. Krueger (1986; see also Lemaire & Fayol, 1995) reported evidence of the use of an odd–even rule of multiplication. Similar to the findings of Campbell and Fugelsang (2001),

this sidestepping strategy is not common in our data, even though, as previously mentioned, this rule was applicable for every false problem in Experiment 2. This outcome suggests that the role of the odd–even rule in verification may have been exaggerated by the reliance on RTs or other performance measures. The data from this study and three previous studies (LeFevre, Sadesky, & Bisanz, 1996; LeFevre, Bisanz, et al. 1996; Siegler, 1987) suggest that any examination of a task that relies completely on external measures (i.e., RTs and errors) runs the risk of exaggerating the importance of some hypothesized strategy.

It is worthy to note, however, that Campbell, Parker, and Doetzel (2004) found that odd–even status did effect RTs for trials in which participants reported using retrieval. Combined with the arguments proposed by Lochy, Seron, Delazer, and Butterworth (2000), this finding suggests that odd–even status may be a property of the representational relationships and not a cognitive strategy per se. These possibilities warrant more investigation before a complete model of mental multiplication can be developed.

Although the present data suggest that problem difficulty and split influence which strategies are used on a specific trial, as suggested by the adaptive strategy choice model (Siegler & Shipley, 1995), these factors are unlikely to be the only ones. Individual-specific factors, such as the participant's knowledge of or ability to use a given strategy and history of success with a strategy, should also help determine which strategy is used. Work by Campbell and colleagues (Campbell & Fugelsang, 2001; Campbell et al., 2004) also suggests that surface features of the problem might also influence strategy choice. Similarly, an $\times 5$ or $\times 9$ rule would not be appropriate for use with a problem that did not have 5 or 9 as an operand. Additionally, the properties of the strategy itself could influence which strategies are available on a given trial. For example, the work of Rickard (1997) suggests that two strategies that entail memory retrieval (e.g., retrieve–compare and calculate–compare) cannot be used concurrently. Rickard's theory does not, however, rule out the possibility that a strategy that entails memory retrieval (either of intermediate results or of a final answer) might be used in parallel with a strategy that does not entail factual memory retrieval (i.e., magnitude estimation strategies).

In summary, the present study provides evidence that arithmetic verification is performed through a mixture of production plus comparison and other sidestepping strategies; researchers' practice of relying solely on performance measures such as latencies or errors may result in conclusions that are incomplete or misleading; and with the right constraints, retrospective reports can be used to gain insight into the processing involved in mental arithmetic without significantly altering normal processing or other basic results.

Appendix A. Verification instructions

In this task, you will be participating in a multiplication verification task. You will be presented with multiplication problems in the form of two numbers and a candidate answer. Your task is to decide if the answer given is true or false for the problem and press the appropriate key as quickly and accurately as possible [go through examples]. Please place one finger of your _____ [left or right] hand on the true key and one finger of your _____ [right or left] hand on the false key. Once we start the experiment please leave your hands in this position until we are finished. After you press the true or false key you will be prompted to report the thoughts you had while doing the problem. At this time, you should report everything you remember from the moment the problem was presented on the screen until you pressed the true or false key. We are interested only in what you actually remember thinking, and everything you can remember is useful to us. Thus, you should report your thoughts in as much detail as possible. You should also report your thoughts in the order in which they actually occurred, from the first thought to the last thought. Think about this as simply playing back a tape of your thoughts from the first thought to the last thought. When you have finished reporting your thoughts the experimenter will tell you to press the enter key to go on to the next trial. Please do not go on to the next trial until you are told to do so. Do you have any questions?

Appendix B. Verification strategies

1. *Retrieve–compare:* Participant reports retrieving the answer, then comparing it with the presented answer. No additional calculation is reported. We will assume that the participant simply retrieved the answer from memory without any intermediate computations.
2. *Calculate–compare:* Participant reports using some (any) intermediate calculation to generate the answer and compares the answer with the presented answer.
3. *Reverse retrieve–compare:* Participant reports thinking of the problem corresponding to the presented answer and then compares the retrieved problem with the presented problem.
4. *Pattern match:* Participant reports simply that the problem just looked true or false, without any intermediate thoughts.
5. *Magnitude estimation:* Participant reports simply knowing that the presented answer could not be correct because the answer was much too large (small).
6. *5 rule:* Participant reports knowing that the presented answer was incorrect (or correct) because there was a mismatch (match) between the $\times 5$ status of the problem and the presented answer.
7. *Odd–even rule:* Participant reports knowing that the presented answer was correct (incorrect) based on the odd–even rule for multiplication.
8. *Explicit no-answer-generation:* Participant explicitly states that he or she did not generate the answer to the problem from memory as a separate step. This should be used any time participants report that they did not know the answer before pressing the "true" or "false" keys, whether or not they report knowing the answer after.

9. *Interference:* Participant reports that the answer first looked correct or incorrect, then he or she realized it was incorrect or correct (perhaps using one of the other strategies).
10. *Switch operands:* Participants report switching operands before using any strategy. This should be coded as the final strategy.
11. *Multiple strategies:* Participant reports using multiple strategies.
12. *Uninterpretable.*
13. *Confusion effects:* Participant reports that a different operation could yield a true verification of the problem presented (i.e., $4 + 4 = 8$, not 4×4; or $8 - 4 = 4$, not 8×4).
14. *9 rules:* Participant explicitly states that he or she used some rule that works only with the 9s table.
15. *Exact square:* Participant reports that he or she knew the answer was either true or false because the answer or operands were an exact square or that the use of any strategy was facilitated by the fact that the operands, answer, or both represented an exact square.
16. *Factor or multiple:* Participant reports that he or she knew that the answer was either true or false because the operands were not factors of the answer or because the given answer was a prime number. Participant reports that the answer was not a multiple of one or both of the operands or that he or she thought of the multiples of one or both of the operands, and the given answer did not match any of them.
17. *Recency effects:* Participant reports that he or she remembered the problem–answer combination or either the problem or answer from the last time he or she saw it and used that information to determine whether the answer was true or false.

Appendix C. Problem set for Experiment 2

Problem	True	False, unrelated, small split	False, unrelated, large split	False related, small split	False, related, large split
7×7	49	48	24	56	28
3×9	27	28	14	24	36
4×6	24	25	45	28	36
3×4	12	14	32	9	21
8×8	64	63	35	72	40
4×7	28	27	54	24	40
5×7	35	36	64	40	20
6×7	42	40	72	36	24
6×6	36	35	64	42	54
4×8	32	30	18	28	20
3×8	24	25	45	27	15
5×8	40	42	21	35	25
3×6	18	16	35	21	27

Continued on next page

Appendix C. *Continued*

Problem	True	False, unrelated, small split	False, unrelated, large split	False related, small split	False, related, large split
3 × 5	15	14	28	12	24
6 × 9	54	56	28	48	72
4 × 9	36	35	15	32	48
7 × 9	63	64	32	56	42
9 × 9	81	64	42	72	54
5 × 9	45	48	24	40	30
5 × 6	30	32	56	35	45
7 × 8	56	54	30	63	35
3 × 7	21	20	36	24	12
6 × 8	48	49	81	54	30
8 × 9	72	63	35	64	48

Notes

Stephen Romero is now at Union College, Schenectady, New York.

This research was supported in part by Army Research Institute contract DASW01-96K-0010 to the University of Colorado. A brief report of this study was presented at the Psychonomic Society's 35th annual meeting.

We would like to thank Alice Healy for her careful and extremely helpful comments throughout this project. We would also like to thank Mark Ashcraft, Jamie Campbell, and Rich Carlson for their useful comments on previous drafts.

Correspondence about this article should be addressed to Stephen Romero, Psychology Department, Bailey Hall Union College, Schenectady, NY 12308 (e-mail: romeros@union.edu).

1. A significant two-way interaction between strategy and the problem difficulty covariate was also found, $F(1, 3) = 243.67$, $MSE = 0.00004$, $p = .01$. Finally, the significant triple interaction between the contrast of true versus false, strategy type contrast, and problem difficulty covariate was also significant, $F(1, 3) = 24.70$, $MSE = 0.00992$, $p = .04$. These effects are included as a footnote for the purpose of completeness but are unpredicted or redundant with other effects reported, and they are not further interpreted.

2. Welford's similarity function is defined as log(larger/[larger − smaller]). Larger values constitute more similarity between the given and correct answer and therefore smaller difference between them, and smaller values as less similarity and therefore larger differences. This function was used because it accounts for effects of the numerical distance between two numbers and for the changes in these effects that occur as numbers become larger (Dehaene, 1989).

3. A significant interaction between the strategy contrast and the problem difficulty covariate was also found, $F(1, 3) = 29.17$, $MSE = 0.0007$, $p = .03$.

References

Ashcraft, M. H. (1982). The development of mental arithmetic: A chronometric approach. *Developmental Review, 2,* 213–236.

Ashcraft, M. H. (1987). Children's knowledge of simple arithmetic: A developmental model and simulation. In J. Bisanz, C. J. Brainerd, & R. Kail (Eds.), *Formal methods in developmental psychology* (pp. 302–338). New York: Springer-Verlag.

Ashcraft, M. H. (1992). Cognitive arithmetic: A review of data and theory. *Cognition, 44,* 75–106.

Ashcraft, M. H. (1995). Cognitive psychology and simple arithmetic: A review and summary of new directions. *Mathematical Cognition, 1,* 3–34.

Ashcraft, M. H., & Battaglia, J. (1978). Cognitive arithmetic: Evidence for retrieval and decision processes in mental addition. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 527–538.

Ashcraft, M. H., & Stazyk, E. H. (1981). Mental addition: A test of three verification models. *Memory & Cognition, 9,* 185–196.

Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology (Vol. 1): Learning, memory & thinking* (pp. 242–293). San Francisco: W.H. Freeman.

Baroody, A. J. (1985). Mastery of basic number combinations: Internalization of relationships or facts? *Journal for Research in Mathematics Education, 16,* 83–98.

Campbell, J. I. D. (1987a). Network interference and mental multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 109–123.

Campbell, J. I. D. (1987b). Production, verification, and priming of multiplication facts. *Memory & Cognition, 15,* 349–364.

Campbell, J. I. D. (1991). Conditions of error priming in number-fact retrieval. *Memory & Cognition, 19,* 197–209.

Campbell, J. I. D., & Clark, J. M. (1992). Cognitive number processing: An encoding complex perspective. In J. I. D. Campbell (Ed.), *The nature and origins of mathematical skills* (pp. 457–492). Amsterdam: Elsevier.

Campbell, J. I. D., & Fugelsang, J. (2001). Strategy choice for arithmetic verification: Effects of numerical surface form. *Cognition, 80,* B21–B30.

Campbell, J. I. D., & Graham, D. J. (1985). Mental multiplication skill: Structure, process and acquisition. *Canadian Journal of Psychology, 39,* 338–366.

Campbell, J. I. D., & Gunter, R. (2002). Calculation, culture and the repeated operand effect. *Cognition, 86,* 71–96.

Campbell, J. I. D., & Oliphant, M. (1992). Representation and retrieval of arithmetic facts: A network-interference model and simulation. In J. I. D. Campbell (Ed.), *The nature and origins of mathematical skills* (pp. 3–39). North Holland: Elsevier.

Campbell, J. I. D., Parker, H. R., & Doetzel, N. L. (2004). Interactive effects of numerical surface form and operand parity in cognitive arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30,* 51–64.

Campbell, J. I. D., & Tarling, D. P. M. (1996). Retrieval processes in arithmetic production and verification. *Memory & Cognition, 24,* 156–172.

Campbell, J. I. D., & Timm, J. C. (2000). Adults' strategy choices for simple addition: Effects of retrieval interference. *Psychonomic Bulletin & Review, 7,* 692–699.

Campbell, J. I. D., & Xue, Z. (2001). Cognitive arithmetic across cultures. *Journal of Experimental Psychology: General, 130,* 299–315.

Dehaene, S. (1989). The psychophysics of numerical comparison: A reexamination of apparently incompatible data. *Perception & Psychophysics, 45,* 557–566.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87,* 215–251.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition, 16,* 309–313.

Hecht, S. A. (1999). Individual solution processes while solving addition and multiplication math facts in adults. *Memory & Cognition, 27,* 1097–1107.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 110,* 306–340.

Kirk, E. P., & Ashcraft, M. H. (2001). Telling stories: The perils and promise of using verbal reports to study math strategies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 157–175.

Koshmider, J. W., & Ashcraft, M. H. (1991). The development of children's mental multiplication skills. *Journal of Experimental Child Psychology, 51,* 53–89.

Krueger, L. E. (1986). Why 2*5 = 5 looks so wrong: On the odd–even rule in product verification. *Memory & Cognition, 14,* 141–149.

LeFevre, J., Bisanz, J., Daley, K. E., Buffone, L., Greenham, S., & Sadesky, G. S. (1996). Multiple routes to solution of single-digit multiplication problems. *Journal of Experimental Psychology: General, 125,* 284–306.

LeFevre, J., & Morris, J. (1999). More on the relation between division and multiplication in simple arithmetic: Evidence for mediation of division solution via multiplication. *Memory & Cognition, 27,* 803–812.

LeFevre, J., Sadesky, G. S., & Bisanz, J. (1996). Selection of procedures in mental addition: Reassessing the problem size effect in adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 216–230.

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

Lemaire, P., & Reder, L. (1999). What affects strategy selection in arithmetic? The example of parity and five effects on product verification. *Memory & Cognition, 22,* 364–382.

Lochy, A., Seron, X., Delazer, M., & Butterworth, B. (2000). The odd–even parity effect in multiplication: Parity rule or familiarity with even numbers? *Memory & Cognition, 28,* 358–365.

Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review, 87,* 252–271.

Richardson-Klavehn, A., & Bjork, R. A. (1988). Measures of memory. *Annual Review of Psychology, 39,* 475–543.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126,* 288–311.

Robinson, K. M. (2001). The validity of verbal reports in children's subtraction. *Journal of Educational Psychology, 93,* 211–222.

Roediger, H. L., & McDermott, K. B. (1993). Implicit memory in normal human subjects. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. 8, pp. 61–131). Amsterdam: Elsevier.

Siegler, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264.

Siegler, R. S., & Shipley, C. (1995). Variation, selection and cognitive change. In G. Halford & T. Simon (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp 31–76). Hillsdale, NJ: Erlbaum.

Stazyk, E. H., Ashcraft, M. H., & Hamann, M. S. (1982). A network approach to simple multiplication. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 320–335.

Stinessen, L. (1985). The influence of verbalization on problem solving. *Scandinavian Journal of Psychology, 26,* 342–347.

Zbrodoff, J., & Logan, G. D. (1990). On the relation between production and verification tasks in the psychology of simple arithmetic. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 83–97.