ARI Research Note 97-33

On Verification of Multiplication Facts: An Investigation Using Retrospective Protocols

Stephen Romero
University of Colorado

Research and Advanced Concepts Office
Michael Drillings, Chief

September 1997

United States Army
Research Institute for the Behavioral and Social Sciences

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency Under the Jurisdiction of the Deputy Chief of Staff for Personnel

EDGAR M. JOHNSON
Director

Research accomplished under contract
for the Department of the Army

University of Colorado

Technical review by

Joseph Psotka

REPORT DOCUMENTATION PAGE

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED (from. . . to)
1997, September	Interim	August 1995-August 1996

4. TITLE AND SUBTITLE	5a. CONTRACT OR GRANT NUMBER
On Verification of Multiplication Facts: An Investigation Using Retrospective Protocols	MDA903-86-K-0010
	5b. PROGRAM ELEMENT NUMBER 0601102A
6. AUTHOR(S)	5c. PROJECT NUMBER B74F
Stephen Romer (University of Colorado)	5d. TASK NUMBER 2901
	5e. WORK UNIT NUMBER C07

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)	8. PERFORMING ORGANIZATION REPORT NUMBER
University of Colorado Department of Psychology Campus Box 345 Boulder, CO 80309-0345	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. MONITOR ACRONYM
U.S. Army Research Institute for the Behavioral and Social Sciences ATTN: PERI-BR 5001 Eisenhower Avenue Alexandria, VA 22333-5600	ARI
	11. MONITOR REPORT NUMBER Research Note 97-33

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES

COR: George Lawton

14. ABSTRACT *(Maximum 200 words)*:

Current theories of mental multiplication elicit two questions: (a) Do the same processes underlie answer production (e.g., 4x7=?) and answer verification (e.g., 4x7=28, T/F?), and (b) Does any theory centered around a single strategy suffice to explain the underlying mechanisms for these tasks? This study involved addition of retrospective protocols to a verification task, in two experiments. The patterns of effects for reaction times (RT) and errors in both experiments were similar to Campbell's (1991) findings, suggesting that the addition of the protocols did not significantly alter the task. Analysis of the protocols provided evidence that retrieval of the correct answer from memory and then comparison to the answer given was the modal strategy reported in both experiments but was not reported for 100% of the trials. These findings imply that the same processes that underlie production are involved. Furthermore, the use of protocols can facilitate differentiating what strategies are involved and provide evidence that any theory of this skill assuming one strategy will likely be incomplete.

15. SUBJECT TERMS

Mental calculation Subject strategies Verbal protocols Memory retrieval

SECURITY CLASSIFICATION OF			19. LIMITATION OF ABSTRACT	20. NUMBER OF PAGES	21. RESPONSIBLE PERSON (Name and Telephone Number)
16. REPORT	17. ABSTRACT	18. THIS PAGE			
Unclassified	Unclassified	Unclassified	Unlimited	63	

ACKNOWLEDGMENT

I would like to thank my committee members Lyle E. Bourne, Jr., Alice F. Healy, and Lewis O. Harvey for their careful comments and support throughout this project. Furthermore, I would like to thank my mother Martha Romero, for without her great support and perseverance I would not have reached this point.

ON VERIFICATION OF MULTIPLICATION FACTS: AN INVESTIGATION USING RETROSPECTIVE PROTOCOLS

CONTENTS

LIST OF TABLES

CONTENTS (Continued)

Page

LIST OF FIGURES

CONTENTS (Continued)

Page

LIST OF FIGURES (Continued)

CHAPTER I

INTRODUCTION

The theories of simple arithmetic are based on data from verification and production tasks. In a production task, subjects are given a problem and asked to either write, type, or say the correct answer. In a verification task, subjects are given a problem and an answer, and are supposed to indicate if the given answer is true or false for the problem given. Much of the discussion about these two tasks revolves around whether the same processes underlie both of them. Further imbedded in this discussion is the question of whether one or more than one strategy underlies verification. In this paper, we explore strategy use in arithmetic verification in detail using retrospective protocol techniques (Ericsson & Simon, 1993).

There are many different models of how people perform verification. Most of the theories assert that subjects use the same method for every problem-answer combination. The question of most importance is, what, then, causes the reaction times to fluctuate from problem to problem? First, Ashcraft and Stazyk (1981) found that for addition problems in a verification format, the numerical distance between the given answer and the true answer was inversely correlated with reaction times. They theorized that magnitude of the difference between the given, incorrect answer and the correct answer provided subjects with a way of bypassing normal processing. Another theory, promoted by Krueger (1986), states that subjects verify whether a given answer is correct by using the odd-even rule for multiplication.

The odd-even rule simply states that if both of the operands of a simple multiplication problem are odd, then the product will turn out odd, and if one or both operands are even the product is also even. Because one or both of the operand's eveness would determine the eveness of the product, subjects should, and were shown to be faster and more accurate in rejecting differences between the given and correct answer of one or three than for differences of two or four. Krueger also reported that subjects did not use the odd-even rule for equations with operands of one or zero. He explains this finding by suggesting that other rules are available to bypass normal odd-even processing.

Campbell's (1987, 1991) findings suggest that the "Retrieve-compare strategy is dominant in adults' arithmetic verification" (Campbell, 1987, p. 350). Campbell argues that verification is a two-stage process, production of the correct answer, and comparison to the answer given. In contrast to other theories, Campbell assumes that the retrieval process is initiated regardless of subjects' intentions. The findings of these studies showed that subjects rejected incorrect answers more slowly than they accepted correct answers, and that incorrect answers that were related to the multiplication table of one of the operands were rejected more slowly than those that were not. In this account Campbell explains differences in reaction times (RT's) in verification as differences in priming effects of correct and incorrect answers. For correct problems, the given answer primes the correct answer therefore facilitating retrieval. For false problems, associative priming from the given answer causes

interference which slows down the retrieval process. A magnification of the interference for table-related problems accounts for the differences in RT's between the two kinds of incorrect answers.

Zbrodoff and Logan (1990) argue that in verification formats subjects compare the whole equation against memory for an earlier instance of the problem and if all the components are the same subjects will respond "correct". This theory assumes that subjects do not calculate in the verification format; only matching of the pattern of the equation as a whole is important. The results reported by Zbrodoff and Logan (1990) were consistent with Campbell's (1987, 1991) data. They explained the differences in reaction times, however, by differential resonance or relative strength of the equations in memory, measured by the frequency of exposure to each equation, resonance being stronger for table-related than table-unrelated incorrect answers.

Clearly, many ambiguities still exist in this area. Campbell (1987) stated that, "The interpretation of verification data depends upon assumptions about the arithmetic processes it incorporates and assumptions about how the verification component interacts with these processes" (p. 349). Similarly, Zbrodoff and Logan (1990) speculated that "it is conceivable that one process may underlie production while another underlies verification" (p. 83). At the basis of this problem is the question of whether one or multiple strategies underlie performance in the verification format.

A common error in theory building is to assume that a single strategy underlies performance on a task that actually reflects several strategies. Siegler's (1987) results illustrate this problem in addition tasks. Siegler argues that by averaging data over different strategies, three factors can lead to incorrect conclusions: (a) relative frequency of each strategy; (b) relative variability of performance generated by each strategy; and (c) independent-dependent variable relations across and within strategies. Relative frequency of each strategy can lead to incorrect conclusions because averaging over a set of frequently used strategies can lead to underestimations of less frequent strategies. The relative variability of performance generated by each strategy as measured by some dependent measure like reaction time can lead to inaccurate conclusions. That is, averaging over strategies with different variances in performance, which are used on equal numbers of trials, allows the strategy with the least amount of variability to have the greatest influence on the pattern of the overall data. Finally, averaging over strategies will not only determine the use of a certain predictor associated with one strategy but will also determine the usefulness of that predictor for other strategies that are not theoretically linked to it. According to Seigler (1987), "When data have been averaged over strategies, the relation between each predictor and the data reflects not only how well the predictor fits the data that is theoretically associated with it, but also how well it predicts data generated by other strategies" (p. 252). Further investigation of

these questions should provide some illumination of theories of arithmetic skill.

What constitutes the most accurate and efficient manner to find out how subjects perform a task? Performance variables (RT's, error rates, etc.) provide only an incomplete picture. Verbal protocols can augment the RT and error data to provide a more detailed picture of arithmetic skills. Past criticisms of verbal reports as data have centered on whether the reports differ from information actually retrieved during performance of the skill. However, growing evidence has shown that following certain guidelines when collecting verbal protocols provides for accurate observations of performance (Ericsson & Simon, 1993).

Use of verbal protocols should help clarify the debate about the verification task, and will help to answer the question whether multiple strategies are used in verifying multiplication facts. Specifically, if the task remains unaltered by the collection of verbal protocols, then the error and RT analyses should replicate and expand some of the findings stated above. Furthermore, if multiple strategies are found to underlie verification, the implications would not only be specific to the theories of arithmetic skill, but could generalize to any theory that assumes the use of only one strategy, a prevalent occurrence in the field.

Chapter II

Experiment 1

Method

Subjects. Twelve subjects from the introductory psychology subject pool participated. All subjects received course credit for their participation.

Apparatus and Materials. Subjects were seated at a table with a computer and tape recording machine in front of them. Subjects were asked to wear a headset microphone adjusted so that a good recording level was held. The experimenter was seated off to one side.

Stimuli were all single-digit multiplication problems. Each problem was presented twice with an incorrect answer and twice with a correct answer. On one of the incorrect presentations, the answer given was table related, and on the other presentation of the problem, the answer given was table unrelated. All of the problem and answer combinations were used in an earlier study by Campbell (1991) in a priming production task.

Design. A 2x3 random block design constituted the framework for this experiment. Both factors, problem size and problem type were inherent in the stimulus set borrowed from Campbell's (1991) study. Problem difficulty was split into two levels, easy and hard, and problem type was defined by three levels, whether the answer given was true, false table unrelated, or false table related. Both problem difficulty and type were variables defined and used in Campbell's (1991) study. Problem difficulty in Campbell's (1991) study constituted a median split

based on the normative RT data from Campbell and Graham's (1985) study.

Procedure. Subjects participated in two sessions lasting one hour each. Three days separated the sessions. Subjects participated one at a time with the same experimenter conducting both of the sessions. In the first session, subjects first were tested in an alphabet verification task, in order to get them comfortable with the way the experimental sessions were conducted. Subjects were given instructions by the experimenter and allowed to ask questions before the computer program was started. Subjects were instructed that a pair of letters would appear on the center of the computer screen and they were 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, subjects were instructed to respond by pressing the key labeled "false". Examples were then presented on a chalk board. Pairs were presented one by one. Subjects were further instructed to place one finger of one hand on the "true" key, and one finger of the other hand on the "false" key, in order to respond as quickly as possible. Both speed and accuracy were stressed in the instructions. After the subjects responded "true" or "false," a prompt for the subjects to remember their thoughts was presented on the screen. At that time, subjects were instructed to report the thoughts they had while working on the

problem from the first moment they saw the problem until they pressed the "true" or "false" key. Subjects were asked to report their thoughts as specifically as possible and in the order in which they actually occurred. After the subjects had reported their thoughts, the experimenter asked for clarification of thoughts that were of interest. After it was clear that the subjects understood the task, the program was started and subjects were presented with 24 trials of the alphabet task. The data from this task were not analyzed. This task was meant to get the subjects used to paying attention to, and reporting their thoughts.

When the subjects were finished with the alphabet task, they were given the instructions for the multiplication task. The instructions were the same except for the fact that multiplication problems were to be presented with candidate answers, and they were to decide whether the given answer was true or false as quickly and as accurately as possible. The block for the multiplication task consisted of 72 trials. Furthermore, the second experimental session was conducted exactly like the first with the exception of the omission of the alphabet verification task from the second session.

Results

The results are presented in three sections, the first two devoted to analysis of the variables addressed in Campbell's (1991) study, reaction time and errors, and the third section presenting the analysis of the protocols. Log base 10 transformations of the reaction times are used instead of raw RT's

in order to reduce any effects due only to heterogeneity of variance or nonnormality of the distributions.

Reaction time data.. Two 3x2 repeated measures analyses of variance, including the correct, incorrect table unrelated, incorrect table related conditions, and two levels of problem difficulty were performed for log RTs, and proportion of errors. The log RT analysis used only those trials for which subjects' responses were the correct response for the given problem-answer combination. The error analysis was performed on the proportion of wrong responses of all trials. The findings of

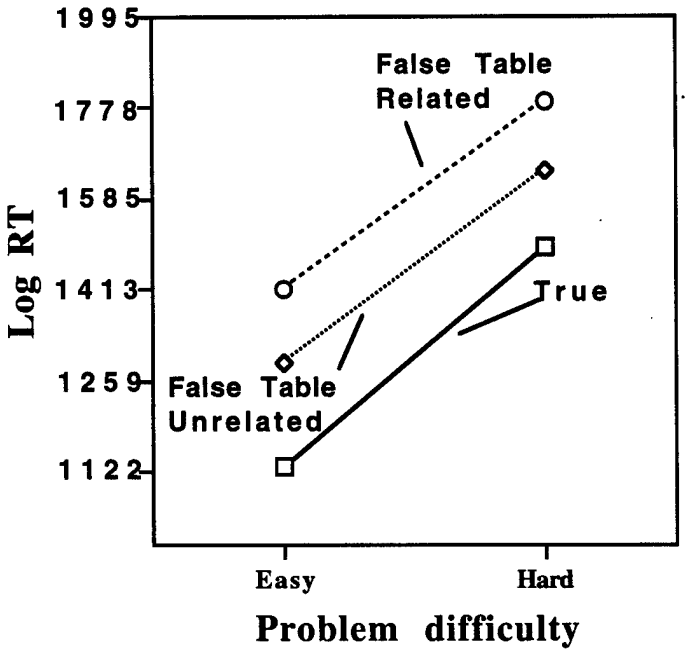

Figure 1. Mean log reaction times for easy and hard problems at levels of problem type.

these analyses were consistent with Campbell's (1991) findings for a primed production task. As shown in Figure 1, the effects of

problem difficulty, $F(1,11)=42.50$ $MSE=.005$, $p<.01$, for log RT's was significant. Specifically subjects were slower, on average, to respond on harder problems than on easy problems. Planned comparisons also showed that subjects were faster, on average, to respond to correct problems than to the average of both types of incorrect problems, $F(1,11)=40.83$, $MSE=.0037$, $p<.01$, for log RT's. Furthermore subjects were faster, on average, to respond to incorrect problems that were table unrelated than to incorrect problems that were table related, $F(1,11)=19.00$, $MSE=.0019$, $p<.01$, for log RT's

Error data.. As shown in Figure 2, the analysis performed on the proportion of errors yielded a significant effect of problem difficulty, such that subjects made more wrong responses, on average, for problems that were hard than for easy problems, $F(1,11)=19.83$, $MSE=.0014$, $p<.01$. Planned comparisons for this analysis also showed that subjects made fewer errors, on average, on problems that were correct than the average of both types of incorrect problems, $F(1,11)=10.57$, $MSE=.0018$, $p<.01$. Furthermore subjects made, on average, more errors on problems where the given answer was table related than on problems with a table unrelated given answer, $F(1,11)=18.38$, $MS=.008$, $p<.01$. Finally a significant interaction was found between problem difficulty and table related and table unrelated answers, such that the differences between the proportion of errors for the table related and table unrelated conditions was greater for hard problems than easy problems, $F(1,11)=15.14$, $MSE=.0014$, $p<.01$.

Protocol analyses. Each trial was categorized into one of 17 different report categories, based on the verbal protocols. Some of the strategy categories were based on a priori theoretical hypotheses and some of the categories were created to group like

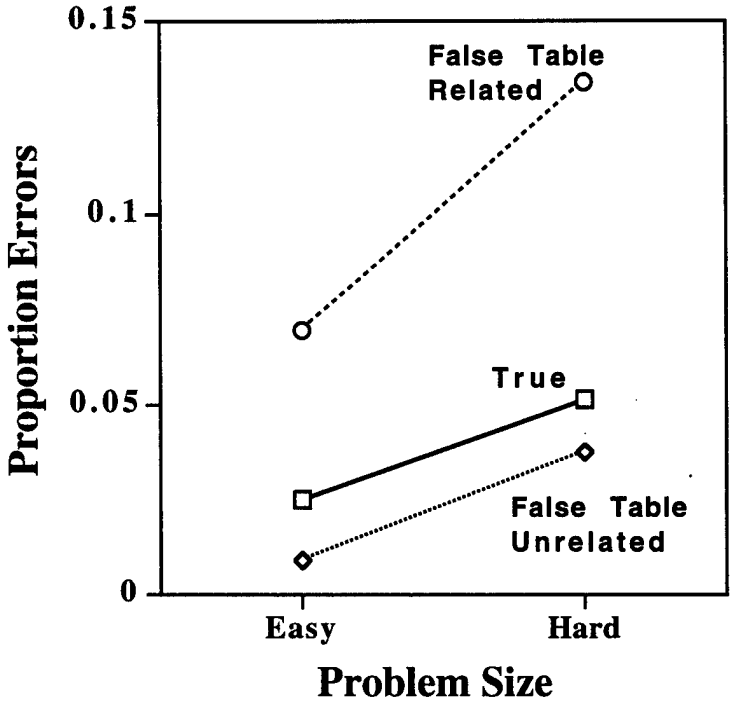

Figure 2. Mean proportion of errors for easy and hard problems at levels of problem type.

protocols together that did not fit into any of the a priori categories. Appendix A presents a description list of all 17 categories. Table 1 presents a breakdown of the most frequent categories the proportion of the overall trials that each strategy was reported and the mean RT for each. If these report categories really represent different strategies, then some specific differences in RTs or some sort of regularity to which problems

they are applied should exist. Teasing out these differences and regularities motivated the following analyses.

Table 1 mean Rt for each report category.

REPORT CATEGORY	Raw Frequency	Proportion of Trials	Mean RT
Retrieve Compare	1154	.67	1442
Calculate Compare	117	.07	2269
Pattern Match	135	.08	1443
Magnitude Estimation	90	.05	1464
Reverse Retrieve Compare	60	.03	1871
Odd-Even rule	4	.002	1006
Other	168	.1	**********

The first of these analysis, shown in figure 3, was a 2x3 repeated measures ANOVA on the calculated proportion of trials in which subjects reported the retrieve compare strategy compared with all the other strategies reported. Trials categorized as retrieve compare were characterized by protocols stating that the subject had produced the correct answer from memory and compared it to the answer given. The two levels of problem difficulty and three levels of problem type that were used in the earlier analyses were used in this analysis. The only significant result was an effect of problem difficulty such that the proportion of trials subjects reported using the retrieve compare strategy compared to all other strategies was smaller for harder problems, $F(1,11)=4.83$, $MSE=.0283$, $p=.05$.

Next a 2x3 repeated measures analysis of covariance was performed for only those trials in which subjects' responses were

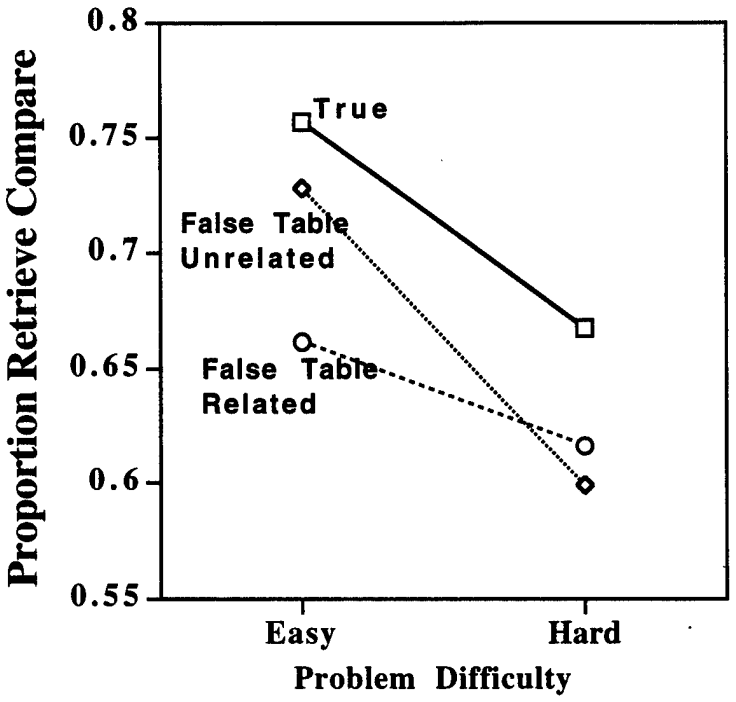

Figure 3. Mean proportion of retrieve compare strategy for easy and hard problems at levels of problem type.

categorized as either retrieve compare or calculate compare. Trials that were categorized as calculate compare were characterized by protocols stating that some intermediate calculating algorithm had been used to produce the correct answer and then it was compared to the answer given. Because using a calculating algorithm implies more processing than 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. The first factor was strategy, either retrieve compare or calculate compare. The second factor was problem

type as defined above. Instead of using the two level, categorical measure of problem difficulty, Campbell and Graham's (1985) normative continuous measure of problem difficulty was used to increase the power of the analysis. Power was a serious concern because only subjects with trials in each cell of the design, four in all, could be used in this analysis. The dependent variable of interest was log RT's. The data are presented in Figure 4.

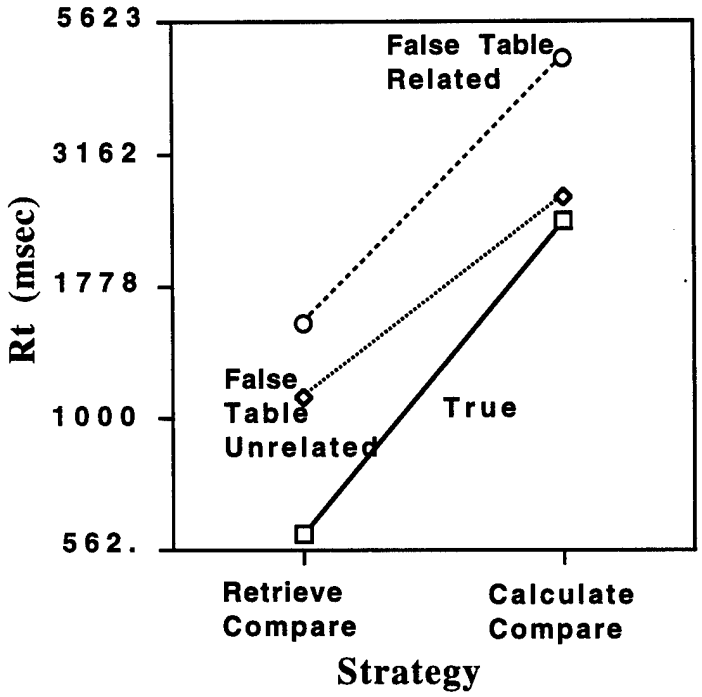

Figure 4. Mean log reaction times for retrieve compare and calculate compare strategies at different levels of problem type.

Planned comparisons with respect to problem type yielded a significant effect of problem type such that subjects' log RT's were faster for true problems than for the average of both types of false problems while controlling for problem difficulty, $F(1,3)=21.73$, $MSE=.00144$, $p=.04$. A significant effect of strategy

was also found such that when subjects reported using calculate compare their log RT's were also slower on average, while controlling for problem difficulty, $F(1,3)=473.02$, $MSE=.00004$, $p<.01$.[1]

The next analyses focused on the magnitude strategy. Magnitude strategy trials were characterized by protocols that stated something to the effect that "The answer was either too large or too small to be right." The first analysis looked for log RT differences between the trials categorized as magnitude and those categorized as retrieve compare, while controlling for problem type, problem difficulty, and the difference between the correct answer and the answer given. Welford's similarity function defined in Campbell and Oliphant (1991)[2] was used as the measure of difference between the given answer and the correct answer. If the magnitude strategy allows the bypassing of normal (retrieve compare) processing, the reaction times should be faster for those trials categorized as magnitude trials, or if the magnitude strategy is used in cases where the correct answer cannot be retrieved, the RTs should be slower for magnitude

---

[1] A significant two way interaction between strategy and a parallel coded problem difficulty contrast was also found $F(1,3)=243.67$, $MSE=.00004$, $p=.01$. Finally, the significant triple interaction between the contrast of true versus false, strategy type contrast, and the parallel coded problem difficulty contrast was found significant, $F(1,3)=24.70$ $MSE=.00992$, $p=.04$. These effects are included as a footnote for the purpose of completeness but are unpredicted or redundant with other effects reported, and are not further interpreted.

[2] Welford's similarity function is defined as the 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.

trials. In either case the ease or frequency of the use of the magnitude strategy should increase for problems with large differences between the given and correct answers. Because the Welford value would be undefined for all correct problems, and the use of the magnitude strategy for correct problems is highly improbable, correct problems were omitted from this analysis. Only the data for subjects who used the magnitude strategy were used for these analyses. The results yielded no significant differences in log RTs. However, in the analysis of

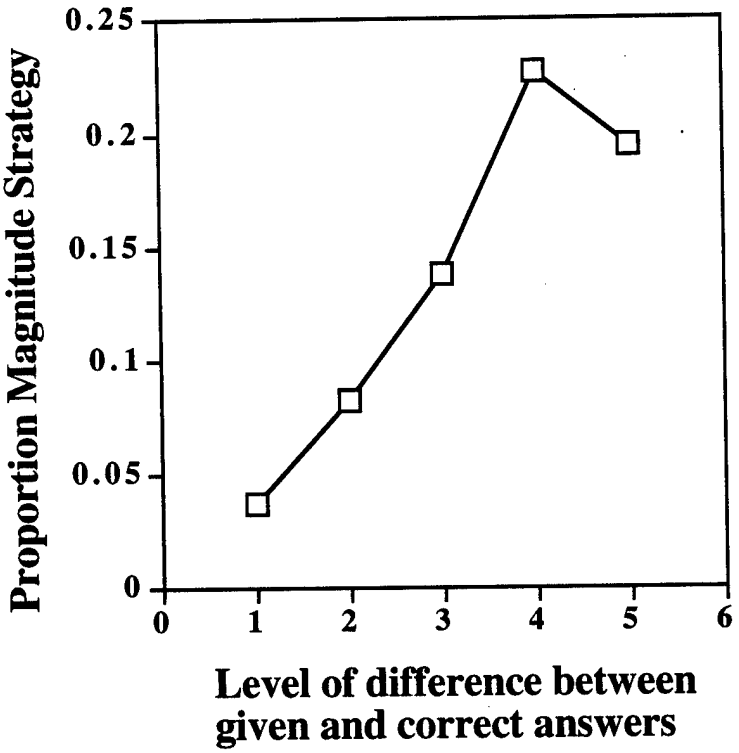

Figure 5. Mean proportion of trials using magnitude strategy at levels of Welford function.

the proportion of trials that subjects reported using the magnitude strategy as a function of Welford values a significant linear trend was found. Specifically as the Welford values increase (increasing similarity between the given and correct answers) the proportion of trials that subjects report using the magnitude strategy decreases, $F(1,8)=9.48$, $MSE=.20303$, $p=.02$. This effect is shown in Figure 5.

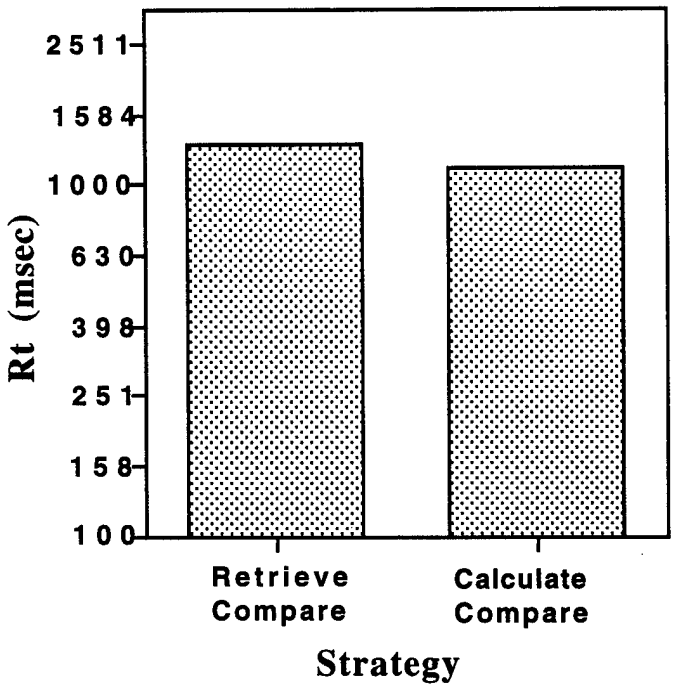

Figure 6. Mean RTs for retrieve compare and pattern match strategies

The final analyses compared the retrieve compare and the pattern match strategies. Trials categorized as pattern match were characterized by protocols stating that the answer just looked right or wrong with no intermediate steps or calculations.

If the pattern match strategy involves no calculation or retrieval, the RTs for trials categorized as pattern match should be faster than those categorized as retrieve compare, and the difference should be more pronounced for true problems. Although this analysis yielded no significant results, an interesting mild trend warrants comment. As seen in Figure 6, the effect of strategy type for true problems approached significance, such that subjects were on average faster on the trials that they reported using pattern match compared to retrieve compare, with problem difficulty controlled, $F(1,5)=3.41$, $MSE=.00086$, $p=.14$. This effect was not evident in the analysis for false problems

Discussion

The results of the overall log RT analysis constitute a direct replication of the patterns of effects in Campbell's (1991) study, even though the two tasks (verification and primed production) are methodologically quite different. The results are also in line with other verification studies (e.g., Koshmider & Ashcraft 1991; Zbrodoff & Logan 1990). Subjects were slower to respond on harder problems and were faster to respond on true problems than on either type of false problem. Subjects were also faster to respond to incorrect problems when the given answer was table unrelated than to those that were table related. The results of the error analysis also constitute an approximate replication of earlier experiments. Subjects made more errors on difficult problems and made fewer errors on correct problems than the average of both types of incorrect problems. Subjects also made more errors on problems for which the given answer was table related than on

those for which the given answer was table unrelated. These replications suggest that protocols did not significantly influence performance in this task. In addition, replications of Campbell's (1991) results help to answer the questions of whether verification and production tasks get at the same processes. Specifically, the similarity of the patterns of the results suggests that the two tasks do have the same underlying processes because Campbell's (1991) study was a priming production task and the present experiment involves a pure verification task.

This analysis, however, cannot account for all the findings of the present study. Evidence was found for the existence of side stepping strategies, like the magnitude strategy, that have no logical link to a production task. This finding would suggest that the real answer to the question of whether the same processes underlie production and verification is, "sometimes." On some of the trials the same processes underlie performance in verification and on some of the trials there are different process in use.

Although Campbell's (1987) assertion that "A retrieve compare strategy is dominant in adults' arithmetic verification," (p. 350) accounts for a large proportion of the trials, it clearly does not account for all the trials, especially when the problems are difficult. Proportion of retrieve compare trials decreases with problem difficulty, from 72% for easy problems to 63% for hard problems. The question of what other strategies account for the remainder of the trials motivated further analyses on the protocol categories to authenticate the categories as actual strategies.

The first of these analyses compared the log RTs of only

those trials that were categorized as either retrieve compare or calculate compare. This analysis showed that, controlling for problem difficulty, subjects were slower when they used the calculate compare strategy than when they used the retrieve compare strategy. One explanation for this finding is that subjects use the calculate compare strategy when they fail to retrieve the correct answer while trying to apply the retrieve compare strategy. In general, calculation can be expected to be slower than retrieval because calculation usually implies more steps to reach a conclusion than one step retrieval (Baroody, 1985). For example, a rule that was commonly reported for problems with nine as one of the operands, was to retrieve the answer to ten times the operand that was not nine and then subtract the non-nine operand from the result. This rule used for problems with a nine as an operand implies two steps to ascertain the correct answer before comparison to the answer given, which should generally take longer than just one retrieval step.

The lack of a significant difference in log RTs between the comparison of the retrieve compare and magnitude strategies would seem, at first glance, to be inconsistent with Ashcraft and Stazyk's (1981) assertion that the size of the difference between the given incorrect answer and the correct answer provides a way of bypassing normal processing, with larger differences being more readily rejectable. If the magnitude strategy is truly used as a way of bypassing normal processing, then the RTs would be expected to be longer or shorter. However, the significant reduction in use of the magnitude strategy with high Welford

values would support their assertion. Since high Welford values signify small differences between the given and correct answers, use of the magnitude strategy would be expected less often with lower Welford values. There is a high level of redundancy between the Welford function and the problem difficulty measures, such that hard problems also had high Welford values and easy problems had low Welford values. This redundancy could possibly mask any RT differences between the strategies. The low power involved with using only four subjects in this analysis could also quite simply account for the inability to detect any significant differences. One other explanation could possibly account for the lack of significant differences: The magnitude strategy could be used as both a backup strategy when retrieval of the correct answer fails, and as a side stepping strategy for bypassing normal processing. If this is the case, the magnitude strategy would produce both faster and slower RTs than retrieve compare and these would wash out in a comparison of strategies.

There were no reliable differences in log RTs between the pattern match and retrieve compare strategies. Possibly, this finding is inconsistent with Zbrodoff and Logan's (1990) notion that subjects compare the whole equation with an earlier instance of the problem. Zbrodoff and Logan concluded that their data ruled out "the possibility that verification is based only on production plus comparison" (p 94). However, they state that their data do not distinguish between verification involving a mixture of production plus comparison and side-stepping strategies or matching the equation as a whole against memory.

The data from the present experiment tend to support the former of these two alternatives with evidence supporting the use of retrieve compare and calculate compare and somewhat supporting the use of the magnitude strategy.

The evidence supporting the uses of other strategies does not rule out the possibility of the use of "resonance" as a side-stepping strategy. However, if the use of "resonance" were then considered a side-stepping strategy and the strategy categories represent what should be quantitatively different strategies (retrieve compare and side-stepping strategies like pattern match) then the possibility of retrieve compare and a side-stepping strategy like pattern match producing RTs that are not reliably different is at best remote. One of two explanations could explain the facts that the log RT differences approach significance for true problems and not for false problems: Pattern matching might only occur for true problems, or pattern matching does not occur at all and these trials are really just instances of fast retrieve compare trials in which subjects are unable to attend to the processes because they occur so fast. If pattern matching only occurs for true problems, then it would follow that when given more power in the reaction time analyses the differences between the strategies should become significant for true problems. If the pattern match trials are really just instances of fast retrieve compare processing, then the increase in power should yield non significant results in all analyses comparing RTs for pattern match and retrieve compare trials.

Experiment 2 attempts to sort out the findings of this study by using a stimulus set that orthoganalizes problem difficulty and Welford values and by reducing the stimulus set to increase the frequency of trials that subjects use the magnitude and pattern match strategies and thereby increase the power of detecting RT differences. Orthoganalizing problem difficulty and Welford values allows for use of the magnitude strategy equally for easy and hard problems which, in turn, will allow for a more balanced comparison of the retrieve compare strategy and the magnitude strategy. Furthermore, by orthoganlizing problem difficulty and Welford values, any influences due specifically to the differences between the given and correct answers or due to problem difficulty should also be apparent in choice and application of other strategies.

Chapter III

Experiment 2

In Experiment 2, an attempt was made to clarify two results from Experiment 1. There were no RT differences between trials on which subjects reported using a magnitude strategy as opposed to using a retrieve compare strategy. However, we did find that, as the similarity between the given and correct answers increased (as measured by Welford's similarity function), use of the magnitude strategy decreased. These two facts seem, at first, contradictory. However, upon further investigation of the stimulus set, we found a high level of redundancy between the problem difficulty and similarity function such that hard problems also had high similarity values and easy problems had low similarity values when presented in their incorrect form. Hence, the question that we are trying to answer in Experiment 2 is whether, if we orthogonalize the difficulty and similarity functions in the stimulus set, we will be able to find RT differences between retrieve compare and magnitude trials and will the pattern of increasing use of the magnitude strategy with larger differences between the given and correct answers replicate when difficulty and magnitude are not confounded.

The second focus of Experiment 2 will be the pattern match strategy. Specifically, we did not find overall RT differences between trials that were categorized as pattern match and those categorized as retrieve compare. As discussed in Experiment 1, a RT difference between retrieve compare and pattern match trials would support the notion that these two categories define

different processing or retrieval mechanisms as defined by Zbrodoff and Logan (1990). In separate analyses for true and false problems the RT differences between trials categorized as retrieve compare and those categorized as pattern match did approach significance for true problems only. The question that remains is whether these are really two different strategies or whether pattern match trials are just very fast retrieve compare trials in which the subjects do not have conscious access to how they performed the verification. If these are not two distinct strategies but instead a single strategy (retrieve compare) then they should occur more often and be faster for easy problems. However, if they are actually distinct strategies then RT effects should persist even if easy problems are removed. For these reasons, we have decided to eliminate some of the easier problems (problems with operands less than 3 and small squares up to and including 5*5) in Experiment 2 and add more subjects to detect reliable differences between these two strategies.

The results of Experiment 2 are expected to provide confirmation of some of the results of Experiment 1 and clarify other results. Significant effects of problem difficulty and problem type congruent with those found in Campbell's (1991) study and in Experiment 1 are expected. First, subjects are expected to take more time to verify and make more errors on problems presented with the average of all types of false primes than true problems. Subjects are also expected to take more time to verify and make more errors on difficult problems than on easy problems. Furthermore, subjects are expected to take more time

to verify and make more errors on problems presented with table related primes than those problems presented with unrelated primes. Finally, regarding the findings of Campbell (1991), the difference in errors for the table related and table unrelated conditions are expected to be more pronounced for hard problems.

Experiment 2 is also expected to replicate the findings of some of the analyses of the protocols from Experiment 1 and clear up results pertaining to the pattern match and magnitude strategies. Subjects are expected to report using the retrieve compare strategy less frequently for harder problems. Subjects are also expected to report using the retrieve compare strategy less for problems presented with high magnitude (i.e., low similarity) answers. Subjects are also expected to be slower to verify problems when they report using the calculate compare strategy than when they report the retrieve compare strategy, and the difference should be more pronounced for harder problems. Use of the magnitude strategy for Experiment 2 should be accompanied by faster verification for problems presented with high magnitude answers than those presented with low magnitude answers. The use of the magnitude strategy should also increase with problem difficulty and the RT differences between problems with high and low magnitude answers should be greater for harder problems. Finally, if the pattern match and retrieve compare strategies constitute separate strategies, then the elimination of the easier problems should not affect the detection of RT differences between the two strategies. However,

the marginal RT effects between these two strategies are not expected to replicate due to the elimination of the easy problems.

Method

Subjects. Sixteen subjects from the introductory psychology subject pool participated. All subjects received course credit for their participation

Apparatus and materials. Subjects were seated at a table with a computer and tape recording machine in front of them. Subjects were asked to wear a headset microphone adjusted so that a good recording level was held. The experimenter was seated to one side.

Problems were all single-digit multiplication problems. Each problem was 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 high similarity to the correct answer. A second was unrelated and of low similarity, a third and fourth were related, and of high and low similarity (where high similarity indicates small differences between the given and correct answers and low similarity indicates large differences between the given and correct answers as indicated by Welford's similarity function). The mean Welford values and their standard deviations are given in Table 2 for each false answer type of problem.

The problem and answer set was based on those used in Experiment 1, with the exception that some easy problems with operands less than three and small squares up to and including

Table 2 Welford values and their standard deviations for each type of problem.

	Table unrelated	Table related
High similarity	Mean=1.32 Standard deviation=.296	Mean=.89 Standard deviation=.088
Low similarity	Mean=.32 Standard deviation=.041	Mean=.46 Standard deviation=.076

5*5 were omitted In addition two more types of answer primes were included to orthogonalize the difficulty and similarity functions used.

Design. A 2x5 random block design was used for this experiment. The first factor, problem difficulty was a median split of problems based on the normative RT data from Campbell and Graham (1985). The second factor, problem type was defined by the properties of the answer prime presented in each trial. Type 1 problems were problems presented with true answers. Type 2 problems containing answers that were table unrelated and of high similarity to the correct answers. Type 3 problems were presented with answers that were table unrelated and of low similarity to the correct answer. Finally Type 4 and Type 5 problems were presented with table related answers that were of high and low similarity, respectively. To help visualize this division, two problems and their answers are presented 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 refer to Apendix B.

Table 3 Example Problems

Problem	False Unrelated High	False Unrelated low	False Related High	False Related low
7x9=63	64	32	56	42
3x9=27	28	14	24	36

Procedure. The procedure for Experiment 2 was the same as that for Experiment 1 with one exception. Subjects in Experiment 2 participated in three sessions. The third session was conducted exactly like the second.

Results

The results are presented in two sections, one focusing on the variables addressed in Campbell's (1991) study and a second section focusing on analyses of the report categories derived from the protocols. Second, log base 10 transformations of the reaction times are used instead of raw RT's to control for any effects due only to heterogeneity of variance or nonnormality of the distributions.

Reaction time data. The first analysis was a 3x2x2 repeated measures ANOVA with log reaction times as the dependent variable. The first factor was problem type that consisted of three levels: true problems, false problems that are presented with answers that are not related to either of the multiplication tables of the operands (table unrelated), and false problems that are

presented with answers that are related to the multiplication table of one of the operands (table related). The second factor is the magnitude of the difference between given false answers and the correct answer (two levels) with high values indicating large differences and low values indicating small differences. The third factor, problem difficulty, also consisted of two levels, easy and hard.

A significant difference in reaction times between problems presented with true answers and the average of all problems presented with false answers was found such that subjects took more time on average to verify false problems than to verify true problems, $F(1,15)=31.16$, $MSE=.0049$, $p<.01$. Second, a significant RT difference between table related and table unrelated problems was found such that on average subjects took longer to verify problems that were presented with answers that were table-related than problems presented with answers that were table-unrelated, $F(1,15)=25.00$, $MSE=.0007$, $p<.01$. Third, a significant effect of problem difficulty was found such that subjects on average took longer to verify hard problems than easy problems, $F(1,15)=21.90$, $MSE=.0084$, $p<.01$. A significant interaction between true and false problems and problem difficulty was also found such that the RT difference between true and the average of all types of false problems is smaller, on average, for hard problems than for easy problems, $F(1,15)=13.99$, $MSE=.0009$, $p<.01$. These four effects are shown in Figure 7. A significant effect of magnitude is shown in Figure 8, such that subjects took

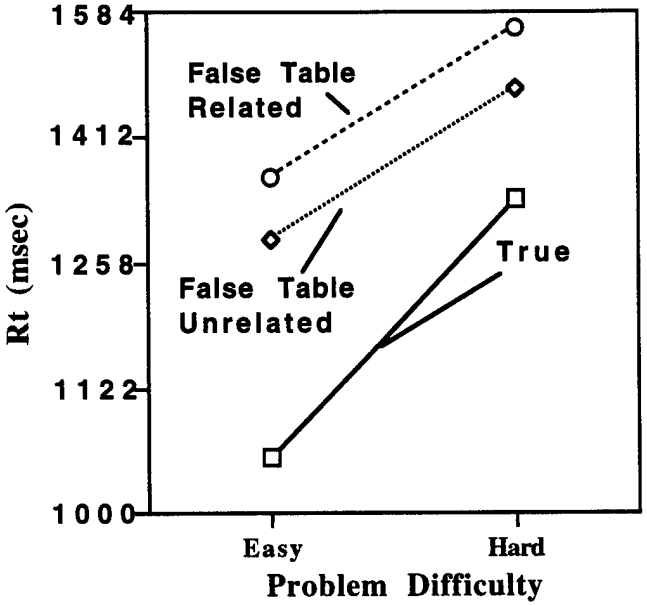

Figure 7. Overall anti-log RT means for True and False problems at levels of problem difficulty.

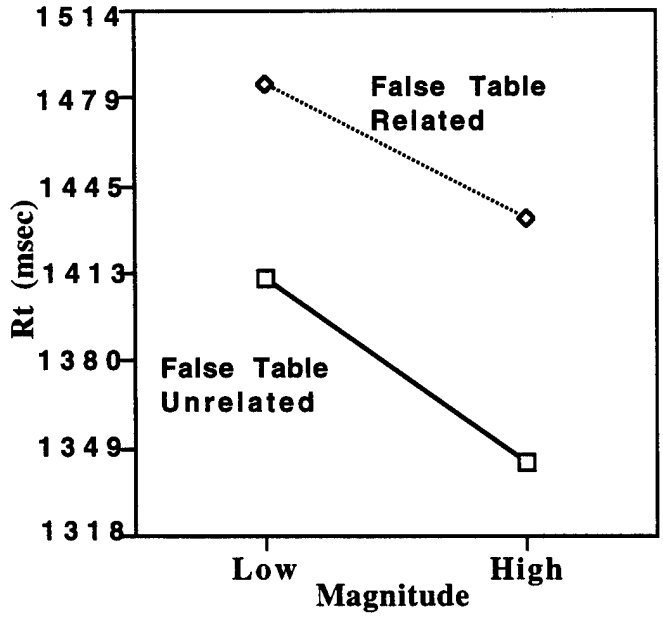

Figure8. Overall anti-log RT means for False problems at levels of magnitude.

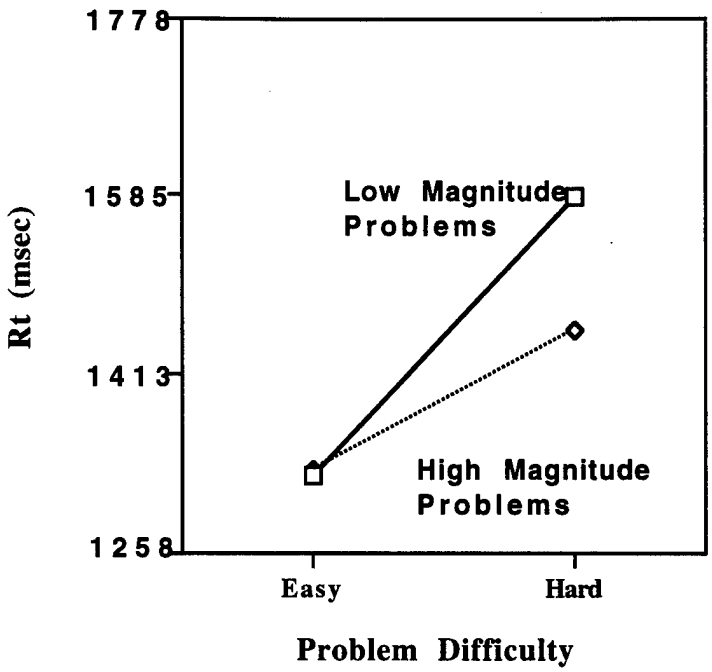

Figure 9. Overall anti-log RT means for low and high magnitude problems at levels of problem difficulty.

more time, on average, to verify problems that were presented with low magnitude answers than high magnitude answers, $F(1,15)=8.02$, $MSE=.001$, $p=.02$. Finally, as shown in Figure 9, a significant interaction of magnitude and problem difficulty was found such that, on average, the RT difference problems presented with low and high magnitude answers was larger for hard problems than easy problems, $F(1,15)=8.81$, $MSE=.001$, $p<.01$.

Errors. The next analysis utilized the same 3x2x2 design as the overall analysis on log reaction times but the proportion of overall errors was the dependent variable of interest. Significant main effects of problem difficulty and magnitude were found in this analysis. As seen in Figure 10, subjects made more errors, on average, on hard problems than on easy problems, $F(1,15)=6.90$,

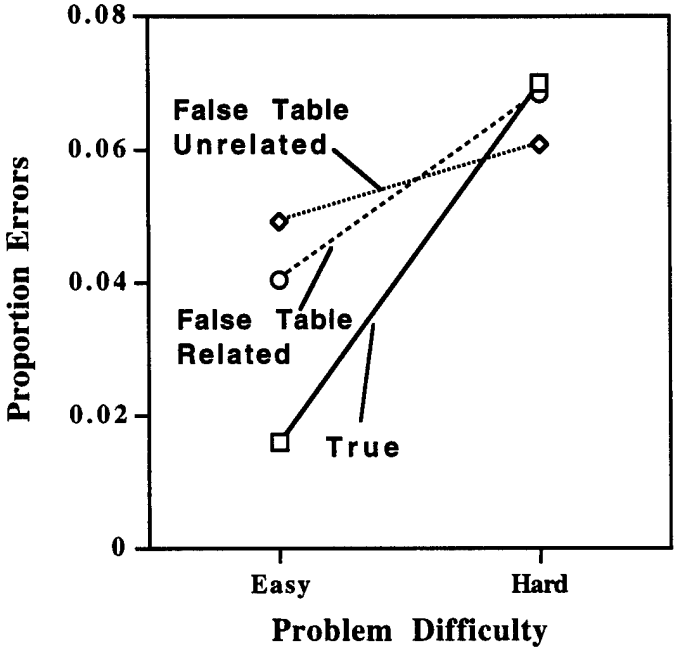

Figure 10. Overall mean proportions of errors for true and false problems at levels of problem difficulty.

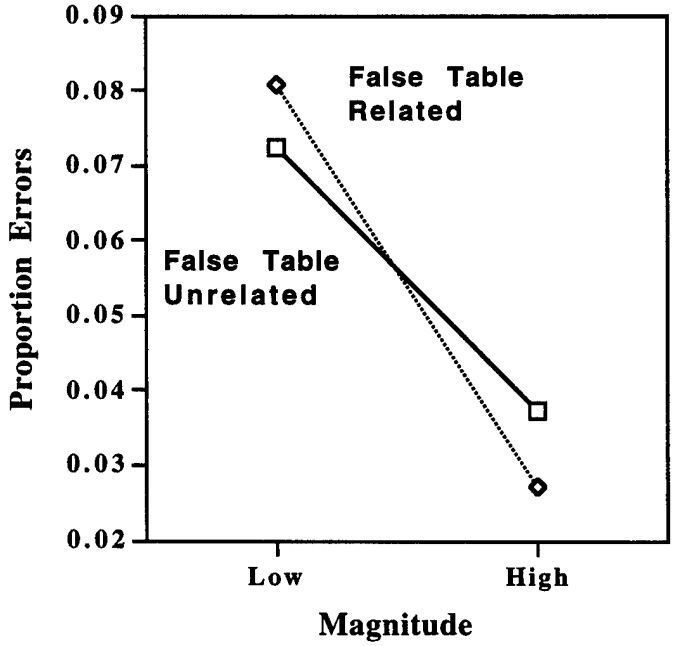

Figure 11. Overall mean proportions of errors for false problems at levels of magnitude.

MSE=.0041, p=.02. As shown in Figure 11, subjects also made more errors, on average, on problems that were presented with low magnitude primes than those presented with high magnitude primes, F(1,15)=8.88, MSE=.0071, p<.01. Predicted differnces in errors between true and false problems as well as differences predicted between table related and table unrelated problems were not found significant.

Protocol Analyses. Each trial in this experiment was categorized into one of 17 different report categories, based on verbal protocols, as in Experiment 1. Some of the report categories were based on a priori theoretical hypotheses and some of the categories were created to group like protocols together that did not fit into any of the a priori categories. A list of all 17 categories can be found in appendix A. A presentation of the frequencies of the occurrence of the most frequently occuring categories, along with the mean RT, and proportion of trials that each strategy is reported for are shown in Table 4.

The following analyses looked for RT differences or patterns of strategy applications that would be consistent with the different strategies identified by the protocols. The first of these protocol analyses investigated trials on which subjects' verbal reports were categorized as retrieve compare. Retrieve compare trials were thoses on which the subject had retrieved the answer from memory and compared in with the answer given. This 3x2x2 ANOVA consisted of three levels of problem type: true, false table related, and false table unrelated; two levels of magnitude: low and high; and two levels of problem difficulty:

easy and hard. The dependent variable was the proportion of trials that subjects reported using the retrieve compare strategy.

Table 4 Report category frequencies proportion of trials and mean RTs

Report Category	Raw Frequency	Proportion of Trials	Mean Rt
Retrieve Compare	1682	.55	1951.
Calculate Compare	96	.03	2256.
Pattern Match	529	.17	1568
Magnitude Estimation	234	.08	1639
Reverse Retrieve Compare	182	.06	1494.
Odd-Even rule	9	.003	1347
Other	321	.1	**********

In replication of Experiment 1, a significant effect of problem difficulty was found such that subjects reported using the retrieve compare strategy less often, on average, for harder problems, $F(1,15)=18.11$, $MSE=.0466$, $p<.01$. Unlike Experiment 1, a significant interaction between true and false problems and problem difficulty was also found such that the difference between true and false problems in the proportion of trials that subjects reported using the retrieve compare strategy was greater, on average, for easy problems than for hard problems, $F(1,15)=8.87$, $MSE=.0116$, $p<.01$. These effects are shown in Figure 12. More interestingly for the purpose of Experiment 2, a significant effect of magnitude was found such that subjects

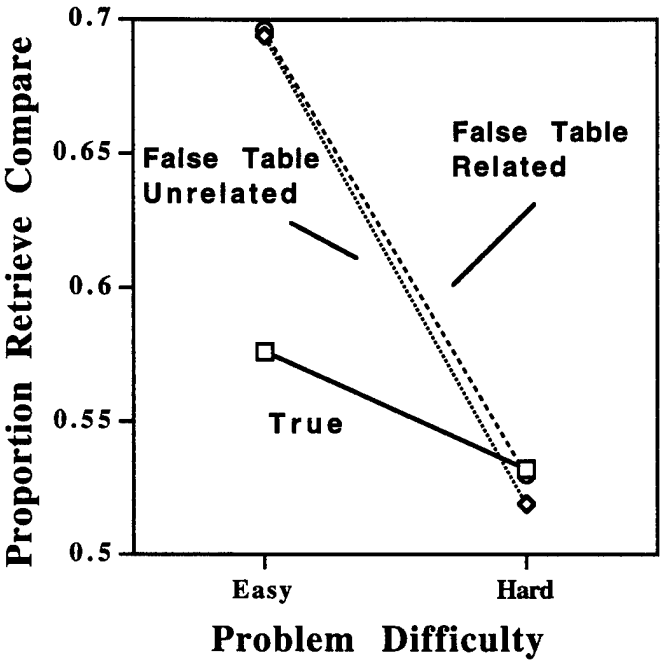

Figure 12. Proportion of trials categorized as retrieve compare for true and false problems at levels of problem difficulty.

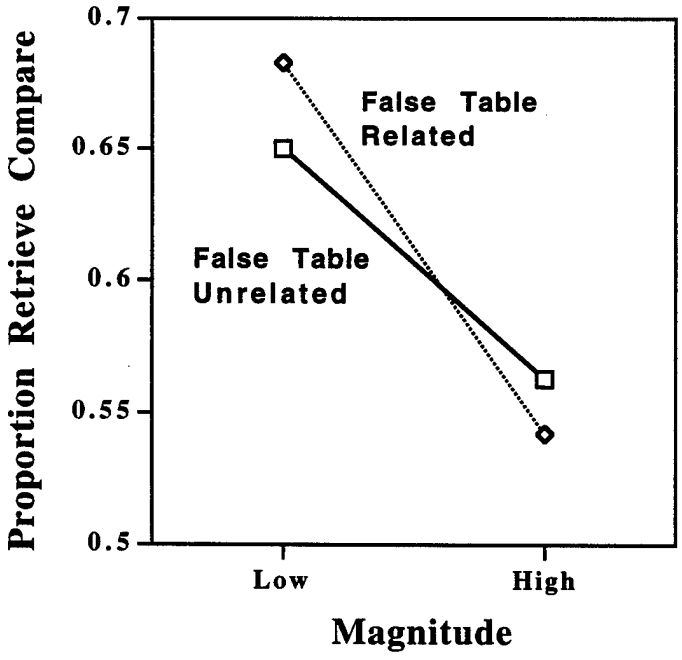

Figure 13. Proportion of trials categorized as retrieve compare for false problems at levels of magnitude.

reported using the retrieve compare strategy less often when the problems were presented with high magnitude answers, $F(1,15)=14.16$, $MSE=.0294$, $p<.01$. This effect is shown in Figure 13.

The next focus was on comparing trials that were categorized as retrieve compare and trials categorized as calculate compare. Trials that were categorized as calculate compare were those on which subjects reported using some calculating algorithm to come up with the correct answer and then compared the result with the answer given. Because the use of a calculating algorithm implies longer processing than retrieval, RTs for trials categorized as calculate compare should be longer than trials categorized as retrieve compare. A 3x2x2 ANCOVA was to be used to investigate these differences. The first factor was problem type, either true, false table related, or false table unrelated. The second factor was strategy, either retrieve compare or calculate compare. The third factor was magnitude either high or low. No subjects had observations in all cells of the design. Due to this problem, this analysis was not possible.

Trials categorized as magnitude strategy were characterized by protocols in which subjects stated that the given answer was either too large or too small to be the correct answer for the problem given. First an analysis was done with the proportion of trials that subjects reported using the magnitude strategy as a function of the difference between the given and correct answers. The level of difference between the given and correct answers

was measured with the use of Welford's similarity function as defined in Campbell and Oliphant (1991). High values for the Welford function correspond with small differences between the given and correct answers and low values correspond with large differences between the given and correct answers. The Welford values were grouped into 5 equally spaced levels with 1 indicating high Welford values (high similarity) and 5 indicating low Welford values (low similarity). Linear through quartic components of trend over Welford values on the proportion of trials that subjects reported using the magnitude strategy were evaluated. Because only subjects that reported using the

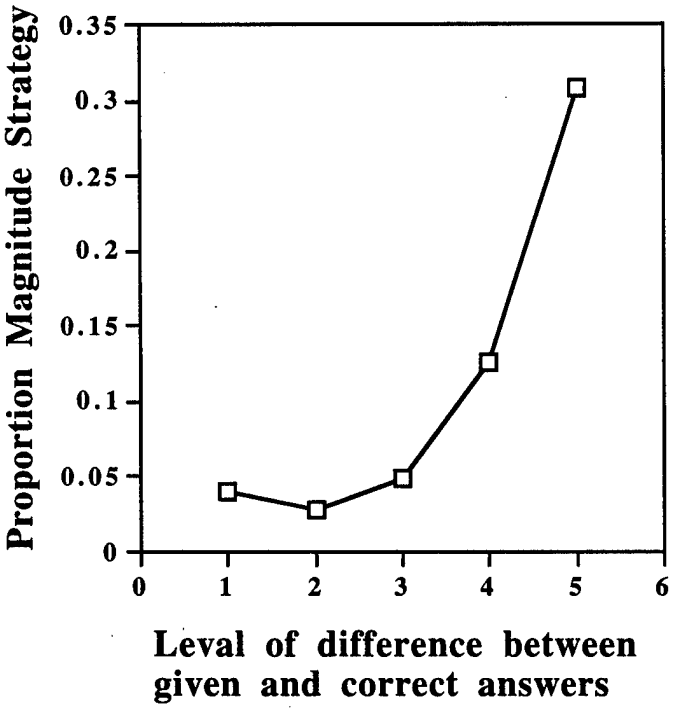

Figure 14. Proportions of trials categorized as magnitude strategy at levels of the difference between the given and correct answers as measured by Welford.

magnitude strategy on 5 or more trials were used for this analysis, only 12 subjects' data were available. As seen in Figure 14, both the linear and quadratic trend components were significant such that as the difference between the given and correct answers becomes larger, indicated by smaller Welford values, subjects report using the magnitude strategy more often, $F(1,11)=38.86$, $MSE=.0124$, $p<.01$ for the linear trend, and $F(1,11)=15.82$, $MSE=.0106$, $p<.01$ for the quadratic trend.

Comparing trials categorized as magnitude strategy and those categorized as retrieve compare was the focus of the next analysis. A 2x2x2 ANCOVA on log reaction times was used for these comparisons. Once again Campbell and Graham's continuous measure of problem difficulty was used as the covariate to increase power. The other factors were two levels of problem type (false table related and false table unrelated), two levels of magnitude and two strategy categories. Only four subjects had observations in each cell of this design and therefore only four could be used in this analysis. True problems were omitted from this analysis because the use of the magnitude strategy is illogical for those problems. In theory the magnitude strategy could be used in two contrasting situations. First, the magnitude strategy could be used to bypass normal (retrieve compare) processing, which would be indicated by faster RTs for the magnitude trials than retrieve compare trials. It is also possible that the magnitude strategy could be used when retrieval is not possible, which would be indicated by slower RTs. This analysis yielded a significant RT difference between trials categorized as retrieve

compare and trials categorized as magnitude such that trials categorized as magnitude were on average faster than retrieve compare trials, with problem difficulty controlled, F(1,3)=27.42, MSE=.0007, p=.03. A significant interaction between strategy and magnitude was also found such that the RT difference between trials categorized as magnitude and trials categorized as retrieve compare was, on average, greater for problems that were presented with high magnitude answers while controlling for problem difficulty, F(1,3)=24.34, MSE=.0004, p=.04. These effects are shown in Figure 15.[4]

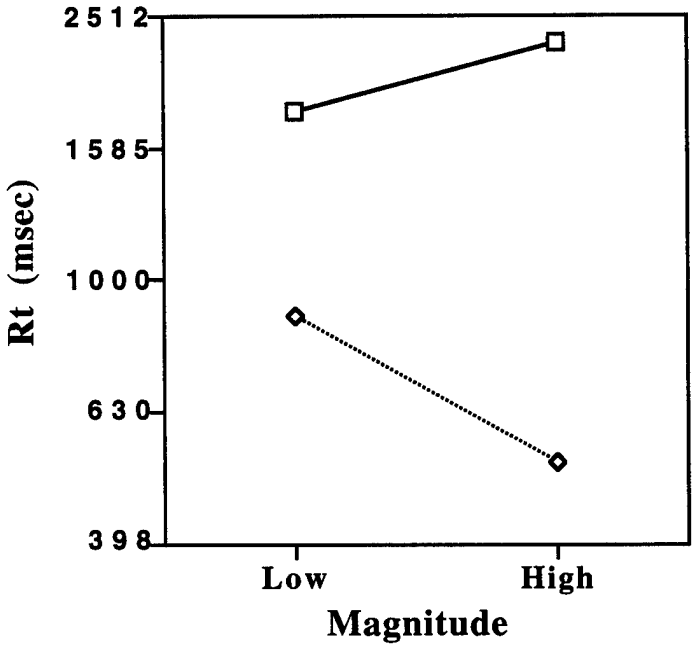

Figure 15. Anti-log mean RTs for trials categorized as retrieve compare and magnitude at levels of magnitude.

[4] A significant interaction between the strategy contrast and the parallel coded dificulty contrast was also found F(1,3)=29.17, MSE=.0007, P=.03 (see footnote 1 for further explanation.)

A comparison of retrieve compare trials with pattern match trials was the focus of the next analyses. Trials categorized as pattern match were characterized by protocols in which subjects reported just knowing the answer was correct or incorrect with no intermediate steps or calculations. A 3x2x2 ANCOVA was conducted with the continuous measure of problem difficulty as the covariate. The three other factors were the three levels of problem type, two levels of magnitude and two levels of strategy: retrieve compare and pattern match. Only six subjects contributed data for this analysis. If the pattern match strategy involves no calculation or retrieval, the RTs for trials categorized as pattern match should be faster than those categorized as retrieve compare and the difference should be more pronounced for true problems. As seen in Figure 16, a marginally significant effect of magnitude was found such that subjects, on average, took less time to verify problems presented with high magnitude answers than low magnitude answers while controlling for problem difficulty, $F(1,5)=5.56$, $MSE=.0011$, $p=.08$. A marginally significant interaction between strategy type and magnitude was also found such that the RT differences between the trials categorized as retrieve compare and trials categorized as pattern match was greater, on average, for problems presented with high magnitude answers when problem difficulty is controlled, $F(1,5)=7.50$, $MSE=.0013$, $p=.05$.[5]

---

[5] A significant difference between true and false problems was also found $F(1,4)=59.03$, $MSE=.0006$, $p<.01$. Finally, a significant interaction between the true Vs false contrast and the parallel coded difficulty contrast was found $F(1,4)=19.59$, $MSE=.0006$, $p=.01$ (see footnote 1 for further explanation.)

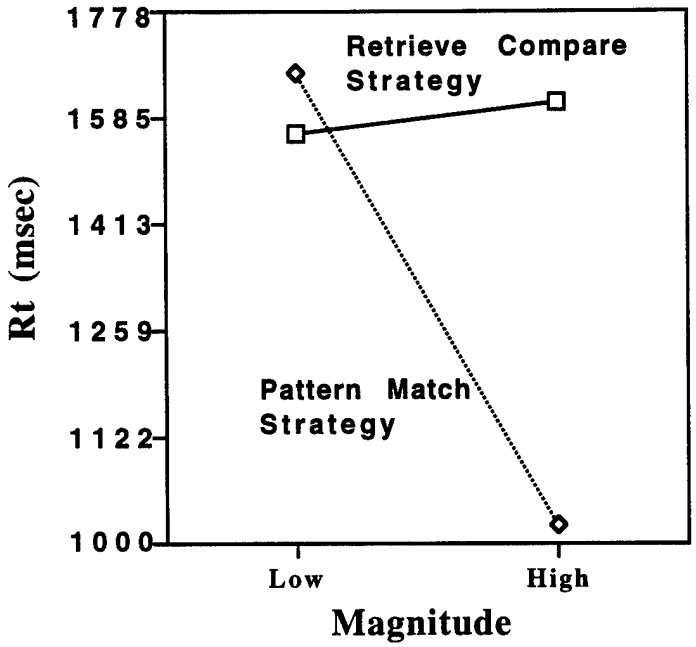

Figure 16. Anti-log mean RTs for trials categorized as retrieve compare and pattern match at levels of magnitude.

Discussion

As was found in Experiment 1, by Campbell (1991) and others (e.g., Zbrodoff & Logan, 1990; Koshmider & Ashcraft, 1991), subjects in Experiment 2 were slower to verify false problems than to verify true problems and slower to verify problems that were presented with table related false answers than table unrelated false answers. Subjects were also slower to verify hard problems than easy problems and the RT difference between true and false problems was smaller for hard problems than for easy problems. Also congruent with the findings of Experiment 1 and earlier studies mentioned above is the finding that subjects in Experiment 2 made significantly more errors on hard problems

than on easy problems. The lack of extensive reliable effects in the error analysis is probably attributable to the fact that errors for Experiment 2 were so near the floor. The replications of these effects from the overall RT and error analyses suggest that the addition of retrospective protocols to the verification task did not alter it in any significant way, which, in turn, lends credibility to the analyses involving the protocols.

The more interesting outcomes from the overall error and RT analyses are the facts that subjects took longer to verify and made more errors on problems that were presented with low magnitude answers than those presented with high magnitude answers. The influences of magnitude 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 task and the factors involved. Zbrodoff and Logan (1990) used the magnitude findings of Ascraft and Stazyk as evidence that production plus comparison is not all that is involved in verification because magnitude effects "suggest that subjects may evaluate the equation as a whole and make their decision without computing or retrieving the true answer. For example, subjects may determine whether the answer is plausible given the arguments" (p. 84). The effects due to magnitude from the present data support this interpretation.

The results of an analysis performed on the proportion of trials that subjects reported using retrieve compare also replicate the findings of Experiment 1 and provide more insight to the influences of magnitude for this task. Subjects reported using the

retrieve compare strategy less often for hard problems than for easy problems and less often for problems that were presented with high magnitude answers. Specifically, retrieve compare was reported for 65% of the easy problems and for 53% of hard problems; whereas retrieve compare was reported for 66% of problems presented with low magnitude answers and 55% for problems presented with high magnitude answers. In conjunction with the influences of magnitude on RTs, the influence of magnitude on the proportion of trials that subjects use retrieve compare also supports the notion that verification involves more than production plus comparison. The evidence that the use of retrieve compare varies depending on the arguments in the equation supports one of the two alternatives proposed by Zbrodoff and Logan (1990) by suggesting that other side-stepping strategies are involved in verification. Furthermore, the findings of this analysis provide insight into subjects' strategy choice in this task. The effects of problem difficulty and magnitude on the proportion of trials that subjects use the retrieve compare strategy further suggest that strategy choice in this task is, in part, influenced by attributes of the problem. Specifically, if the difference between a given false answer and the correct answer is large, then subjects are more likley to use choose a strategy that uses this information for verification (e.g., the magnitude strategy). If the difference between the given and correct answers was small, subjects' choice of a strategy that uses magnitude would be less likely and subjects would be more likely to choose retrieve compare, calculate compare or some other

strategy that uses some other information contained in the problem (e.g., a x5 rule, a x9 rule, if 9 or 5 were one of the operands). On problems that are presented with low magnitude answers subjects may also choose production-like strategies such as retrieval or calculate compare. However, because subjects reported using retrieve compare less often for hard problems when hard problems are presented with low magnitude answers the use of retrieval is less likely than the use of either calculate compare or some side-stepping strategy like the x5 or x9 rules. Problem difficulty and magnitude are probably not the only factors that influence strategy choice in this task. The manner in which subjects were taught the numerical relationships involved in the verification task will most certainly have effects on strategy choice, as will individual differences such as an individual's history of success with the strategy. Factors such as these need to be investigated further and possibly included in theories of mental calculation.

The most interesting, and most readily interpretable results of Experiment 2 involve trials on which subjects reported using the magnitude strategy. In replication of Experiment 1, the prediction that subjects would use the magnitude strategy more often as the difference between the given and correct answer got larger was realized. What is more important, when subjects reported using the magnitude strategy, they verified problems faster than when they reported using retrieve compare and the RT differences were greater for problems that were presented with high magnitude answers. It follows from the findings

concerning the magnitude strategy, that subjects 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. Again, the existence of side stepping strategies supports the notion that verification involves both production plus comparison and side stepping operations.

The notion that pattern match and retrieve compare trials represent quantitatively different strategies or retrieval mechanisms was not supported by the data of Experiment 2. As predicted, the marginal RT effects that were found for true problems between the pattern match and retrieve compare trials did not replicate in Experiment 2, nor was any significant main effect due to strategy found. When verbal reports were categorized as pattern match subjects were marginally faster to verify problems that were presented with high magnitude answers than when their verbal reports were categorized as retrieve compare; but there was no difference when the problems were presented with low magnitude answers. The interaction between strategy and magnitude is not necessarily consistent with the matching the equation as a whole against memory. If trials that were categorized as pattern match are really just fast retrieve compare trials that subjects do not have conscious access to, then the interaction between strategy and magnitude can be explained as a function of the comparison stage of retrieve compare. Specifically, larger differences between the given and correct answers would facilitate functioning at the comparison

stage. This explanation is consistent with Ashcraft and Stazyk's (1981) findings and with other with work done under the rubric of visual perception by Johnson (1939) who asked subjects to judge the length of lines, and later by Moyer (1973) who had subjects judge the relative size of animals in memory. In both cases the time for subjects to make the comparison was a function of the difference of the stimuli presented with larger differences being faster.

Chapter IV

General Discussion

Retrospective verbal protocols have proven useful in furthering knowledge of the processes and mechanisms that are involved in mental calculation. In particular, the present study has provided evidence that should help to clarify several questions about the relationship between tasks commonly used in investigations of mental calculation. Replication of the patterns of effects found in Campbell's (1991) study supports the assumption that protocols do not influence in any way the basic processes involved in mental calculation. Campbell used a primed production task, and the present study used a pure verification task. The similarity of the patterns of effects from two different tasks further supports the assertion that the same processing involved in production underlies performance in a large proportion of verification trials but production processes do not account for everything involved in verification. The previous statement is in general agreement with one of the conclusions of Zbrodoff and Logan (1990); that is, that production plus comparison are not all that is involved in verification. However, the present data do not agree with all the conclusions of their study. Zbrodoff and Logan (1990) argue that verification involves either production plus comparison with the use of other side stepping strategies or a comparison of the equation as a whole against memory. They argue further for the option that subjects compare the equation as a whole against memory and thereby retrieve a measure of degree of match, and against the

option that verification involves production plus comparison with the use of other side-stepping strategies. Zbrodoff and Logan (1990) argue that the retieval of a measure of degree of match will either exceed some threshold and produce a true response or will not exceed the threshold and thereby produce a false response. Data from the present experiments support the assertion that verification involves production plus comparison with the use of side stepping strategies. Specifically, in the present study, the explicit reports from the subjects stating that they either retrieved the answer from memory or calculated the correct answer by some algorithm that can also be explicitly reproduced supports the use of production-like processing in this task. The agreement of performance measures with the use of retrieval and calculation further supports the use of production-like processing in verification. Moreover, the explicit reports of side-stepping strategies and the agreement of performance measures with these reports imply the use of non-production side-stepping processing in verification. Therefore the evidence supporting both production-like processing and non-production side-stepping processing supports the assertion that verification involves production plus comparison and side stepping strategies, an assertion that has important implications for models of mental calculation.

Is it possible that the use of resonance constitutes different processing that should be included as a side-stepping strategy in verification? The present data do not support the occurrence of resonance in verification. The lack of reaction time difference

between the pattern match and retrieve compare strategies is not consistent with what should be quantitatively different processing. Instead the data seem to support the assertion that pattern match trials are actually instances of retrieve compare trials in which processing is too fast for any cues for reconstruction of the processes involved to be available for subjects' reports. The lack of evidence for the use of resonance in verification would suggest that mental calculation models that revolve around the use of resonance are not valid.

What are the implications of the existence of side-stepping strategies for non-resonance models of mental calculation? The present data quantitatively supports the use of the magnitude strategy. However, the explicit reports of many other candidates of side-stepping strategies could not be analyzed due to low frequencies of use (an inspection of Appendix A will yield a general description of other candidate side-stepping strategies). Some of these strategies have been investigated in the literature on mental calculation (e.g., Krueger, 1986; Lemaire & Fayol 1995)[6] and some have not. Although the present data suggest that problem difficulty and magnitude influence which strategies are used on a specific trial, these are probably not the only factors involved. Other surface features of the problem might also influence strategy choice. A x5 or x9 rule would not be appropriate for use with a problem that did not have 5 or 9 as an

---

[6] Both studies cited here present evidence for the odd-even rule of multiplication. The present data, however, suggests (see tables 1 & 4) that the use of this of this side-stepping strategy is not as wide spread as these authors suggest.

operand. Individual specific factors like the individual's knowledge of, skill or ability to use a given strategy and history of success with a strategy could also help determine which strategy is used. Specific individual differences factors could prove difficult for a general model of the task. Each subject does not enter into the task with the same distribution of strategies. Furthermore subjects might not have the same distribution of strategies at two different stages in the development of the mental calculation skill. Any general theory or model of mental calculation must account for the differing distributions of strategies. Otherwise for reasons that Siegler (1987) pointed out, any general theory that revolves around one strategy or one distribution of strategies will likely be incomplete.

References

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

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

Campbell, J. I. D. (1987). 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., & Graham, D. J. (1985). Mental multiplication skill: Structure, process and acquisition. Canadian Journal of Psychology, 39, 338-366.

Campbell, J. I. D., & Oliphant, M. (1991). Representation and retrieval of arithmetic facts a network-interference model and simulation. in J. D. Campbell The nature and origins of mathematical skills (pp. 3-39). North Holland: Elsevier Science Publishers.

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

Johnson, D. M. (1939). Confidence and speed in the two-category judgment. Archives of Psychology, 241, 1-52.

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

Koshmider, J. W., & Ashcraft, M. H. (1991). The development of children's mental multiplication skills. Journal of Experimental Child Psychology, 51(1), 53-89.

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.

Moyer, R. S. (1973). Comparing objects in memory: Evidence suggesting an internal pschophysics. Perception and Psychophysics, 13, 180-184.

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.

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

Appendix A

Verification Strategy List

The following numerical values should be entered in the "strat" column of the data file as appropriate:

1) Retrieve-Compare: Subject reports retrieving the answer, then comparing it with the presented answer. No additional calculation is reported. We will assume that the subject simply "retrieved" the answer from memory, without any intermediate computations.

2) Calculate-Compare: Subject reports using some (any) intermediate calculation to generate the answer, and compares the answer with the presented answer.

3) Reverse Retrieve-Compare: Subjects reports thinking of the problem corresponding to the presented answer, and then compares the retrieved problem with the presented problem.

4) Pattern Match: Subject reports simply knowing the answer was true or false without any intermediate thoughts.

5) Magnitude Estimation: Subject reports simply knowing that the presented answer could not be correct becasue the answer was much too large (small).

6) x 5 rule: Subject reports knowing that the presented answer was incorrect (or correct) because there was a mismatch (match) between the x5 status of the problem and the presented answer.

7) Odd-Even rule: Subject reports knowing that the presented answer was correct (incorrect) based on the odd-even rule for multiplication.

8) Explicit no-answer-generation: Subject explicitly states that he did not generate the answer to the problem from memory as a separate step. This should be used anytime subjects' 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: Subject reports that the answer first looked correct or incorrect, then they realized it was incorrect or correct (perhaps using on of the other strategies). [This will be worthwhile because, if it occurs with any frequency, we can then evaluate whether associatively related candidate answers yield more of this than unrelated answers.]

10) Switch opperands Subjects report switching operands before using any strategy. this should be coded as the final strategy.

12) Uninterpretable.

13) Confusion Effects: Subject 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: Subject explicitly states that they used some rule that only works with the nines table.

15) Exact Square: Subjects report that they either knew the answer was true or false because either 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 represents an exact square.

16) Factor or Multiple Subject reports that they knew that the answer was either true or false because the operands were not factors of the answer or that the given answer was a prime answer. Subjects report that the answer was not a multiple of one or both of the operands or that they thought of the multiples of one or both of the operands and the given answer did not match any of them.

17) Recency Effects: Subjects report that they remembered the problem answer combination or either the problem or answer from the last time they saw it and used that information to determine whether the answer was true or false.

Appendix B

Problem set for Experiment 2

Prob.	True	False Unrelated High	False Unrelated low	False Related High	False Related low
7x7	49	48	24	56	28
3x9	27	28	14	24	36
4x6	24	25	45	28	36
3x4	12	14	32	9	21
8x8	64	63	35	72	40
4x7	28	27	54	24	40
5x7	35	36	64	40	20
6x7	42	40	72	36	24
6x6	36	35	64	42	54
4x8	32	30	18	28	20
3x8	24	25	45	27	15
5x8	40	42	21	35	25
3x6	18	16	35	21	27
3x5	15	14	28	12	24
6x9	54	56	28	48	72
4x9	36	35	15	32	48
7x9	63	64	32	56	42
9x9	81	64	42	72	54
5x9	45	48	24	40	30
5x6	30	32	56	35	45
7x8	56	54	30	63	35
3x7	21	20	36	24	12
6x8	48	49	81	54	30
8x9	72	63	35	64	48