*Memory & Cognition*
*2000, 28 (1), 1-7*

# When it hurts to be misled: A Stroop-like effect in a simple addition production task

N. JANE ZBRODOFF and GORDON D. LOGAN
*University of Illinois at Urbana–Champaign, Champaign, Illinois*

In four experiments, subjects saw simple addition equations (e.g., 3 + 4 = 9) and produced the sums while ignoring the presented answer. If the presented answer was false, subjects took longer to produce the sum, as compared with when the presented answer was true (Experiment 1), when there was no answer presented (blanks; Experiment 2), when a letter was presented (Experiment 3), and when a symbol was presented (Experiment 4). The results suggest that subjects were unable to ignore the presented answers, which raises problems for theories of arithmetic verification (i.e., deciding whether 3 + 4 = 9 is true or false) that claim that subjects verify equations by first producing the sum and then comparing the produced sum with the presented answer. Our results are more compatible with theories that claim that in verification and production, an arithmetic knowledge base is used in different ways.

The psychology of simple arithmetic is based primarily on two main tasks, *production* and *verification*. In production tasks, subjects are presented with a problem (e.g., 5 + 2 =) and are asked to produce its answer. In verification tasks, subjects are presented with a problem and an answer (e.g., 5 + 2 = 7), and they are asked to say whether the answer is true or false. This article concerns the relation between production and verification.

One hypothesis about the relation, perhaps the first in the field, is that verification includes production (Groen & Parkman, 1972; Parkman, 1972; Parkman & Groen, 1971). Subjects perform the verification task by producing an answer and then comparing it with the presented answer, saying "true" if it matches and "false" if it does not. Thus, verification is production plus comparison. An alternative hypothesis is that verification and production operate differently because they tap the knowledge base of arithmetic facts in different ways (Zbrodoff & Logan, 1990). Verification is analogous to recognition in standard memory tasks. Subjects use the whole equation—the digits and the putative sum or product—as a retrieval cue, and they evaluate the activation or resonance that results from the retrieval cue. The decision is based on overall activation; true problems produce more activation than false ones. Production is analogous to recall in standard memory tasks. Subjects use the digit arguments as retrieval cues and select the most active item in memory as the answer. The decision is based on relative activation; the true answer is more active than the alternatives, and it can be selected using something like Luce's (1963) choice rule (see, e.g., Gillund & Shiffrin, 1984; Murdock, 1993).

Both hypotheses receive support in the literature. The idea that verification is production plus comparison predicts that factors like the magnitude of the digit arguments that affect the production process should have the same effect in production and verification tasks. Sometimes argument magnitude has the same effect in production and verification (Ashcraft, Fierman, & Bartolotta, 1984), and sometimes it does not (Campbell, 1987; Campbell & Tarling, 1996).

The hypothesis that verification is production plus comparison also predicts, following Sternberg's (1969) additive factors logic, that factors that affect the production process will not interact with factors that affect the comparison process, such as the difference between true and false answers. The evidence here is also mixed. On the one hand, Groen and Parkman (1972; Parkman & Groen, 1971) found no interactions between argument magnitude and the difference between true and false answers, and Geary, Widaman, and Little (1986) found no interaction between argument magnitude and *split* (i.e., the difference between "near" and "far" false answers—e.g., 3 + 4 = 9 vs. 3 + 4 = 19). On the other hand, Ashcraft and Stazyk (1981); Stazyk, Ashcraft, and Hamann (1982); and Campbell (1987) found interactions between argument magnitude and split.

These results are hard to reconcile with the hypothesis that verification is production plus comparison. It predicts equal argument magnitude effects and null interactions between argument magnitude and split. However, it is not clear that these results are consistent with the hypothesis that verification and production access the same knowledge base in different ways. That hypothesis does not have to predict equal argument magnitude effects and null interactions, but, without a formal model, it is not clear what it does and does not predict. It may be better to try to distinguish the hypotheses in a manner that does not depend on specific models.

This research was supported by National Science Foundation Grants SBR 9709711 and SBR 9808971. We are grateful to Julie Delheimer for testing the subjects and to Jamie Campbell and Lester Krueger for helpful comments on the manuscript. Correspondence concerning this article should be addressed to N. J. Zbrodoff, Department of Psychology, University of Illinois, 603 East Daniel Street, Champaign, IL 61820 (e-mail: jzbrodof@s.psych.uiuc.edu).

The hypothesis that verification is production plus comparison has difficulty explaining several demonstrations that subjects may evaluate the equation as a whole and decide on their response without producing the true answer. Ashcraft and Stazyk (1981) showed that subjects determined whether answers were plausible, given the arguments, and rejected problems with extreme splits very quickly. Zbrodoff and Logan (1986) showed that subjects were slower when a false answer was true for another operation (e.g., 3 + 4 = 12) than when it was not (e.g., 3 + 4 = 11). Krueger and Hallford (1984; also see Krueger, 1986) showed that subjects rejected false answers quickly when they violated parity rules (in addition, the answer is even if both arguments are even or both are odd; the answer is odd if one argument is even and the other is odd). These results suggest that during verification tasks subjects may either sidestep the production process entirely or simply evaluate the extent to which the arguments and answers together resonate with memory.

Other results challenge the hypothesis that verification is production plus comparison. Campbell and Tarling (1996) had subjects alternate between production and verification and found that *error priming* was much stronger within tasks than between tasks. That is, the errors that subjects made tended to be repetitions of correct answers to problems presented 5–20 trials earlier. However, errors on the production task were more likely to be correct answers from previous production trials than from previous verification trials, suggesting that production and verification processes are quite different. Dagenbach and McCloskey (1992) found a brain-damaged patient whose impairment in production was greater than his impairment in verification. If verification was production plus comparison, verification could not have been better than production.

Zbrodoff and Logan (1990) tried to bridge the procedural gap between verification and production by presenting the arguments and the operation symbol (e.g., 3 + 4 =) in one display and the whole equation (e.g., 3 + 4 = 7) in another, varying the delay between displays. The subjects were asked to verify the equation in the second display. If verification were production plus comparison, at long delays there should be no argument magnitude effect. The subjects should have produced the true answer during the delay, so they could compare it directly with the presented answer in the second display. The data did not confirm this hypothesis. There was a robust argument magnitude effect in reaction times to the second display, even when the subjects were allowed to control the delay between displays so that they would have time to produce the answer themselves. Zbrodoff and Logan (1990) concluded that subjects waited for the whole equation to appear and then based their verification decisions on its resonance with memory. Thus, verification was not production plus comparison; there was no production in verification.

## THE PRESENT EXPERIMENTS

Our approach to the relation between production and verification is that production-plus-comparison and resonance evaluation are different strategies for performing verification tasks. The evidence reviewed above suggests that subjects can use the resonance evaluation strategy, and a moment's introspection will reveal that people can use the production-plus-comparison strategy if they wish to (try it yourself on 9 + 7 = 15). Thus, we are concerned more with the prevalence of each strategy in the verification task than with deciding whether verification should be modeled in one way and not the other. The utility of alternative verification strategies depends on the specific experimental conditions under which verification is attempted. We were inspired by a series of experiments by Campbell and his colleagues (Campbell, 1987, 1991; Campbell & Tarling, 1996; Meagher & Campbell, 1995) that suggested that the answer in a verification task may prime retrieval of the correct answer. If such priming occurs in the standard verification procedure, subjects may choose to rely on resonance evaluation rather than production plus comparison to verify answers.

Campbell and his colleagues (Campbell, 1987, 1991; Campbell & Tarling, 1996; Meagher & Campbell, 1995) had subjects perform a production task, in which correct answers, false answers, and neutral symbols were presented as primes 200–1,700 msec before the production task stimulus. They found that true answer primes facilitated production, relative to neutral primes, whereas false answer primes inhibited production, relative to neutral primes. Meagher and Campbell argued that two processes were responsible for these priming effects. One was a fast-acting automatic process that affected retrieval of arithmetic facts at short stimulus onset asynchronies (SOAs), and the other was a slower strategic process that operated at longer SOAs, in which subjects reported the prime if the prime and problem together were sufficiently familiar.

We were interested in the automatic priming effect because of its implications for the effectiveness of production-plus-comparison and resonance evaluation strategies in verification tasks. The procedure of the verification task invites priming. A true or false answer is always presented with the arguments, and that presented answer may prime retrieval of the correct answer (Campbell, 1987, 1991; Campbell & Tarling, 1996; Meagher & Campbell, 1995). This priming may have been detrimental for subjects who chose to verify with the production-plus-comparison strategy. Priming would add noise to the retrieval process, making it more difficult to choose the correct answer (Campbell, 1991). By contrast, subjects who relied on resonance would benefit from priming, since it adds to the difference in activation between true and false problems (Campbell, 1991; Zbrodoff & Logan, 1990).

In our experiments, we required subjects to simulate a portion of the production-plus-comparison strategy for verification by requiring them to produce sums of digits presented in the format of a verification problem. The subjects saw two digit arguments and a true or false answer in the answer field (e.g., 3 + 4 = 9), just as they would in a verification task. However, their task was to produce the sum of the digit arguments while ignoring the presented answer, which is essentially the first step in the production-plus-

comparison verification strategy. We assessed priming by comparing reaction times (RTs) to problems with false answers with RTs to problems with true answers in Experiment 1 and with problems with various control stimuli in the answer field in Experiments 2–4. Our goal was to determine whether subjects attempting the production-plus-comparison strategy in a verification task would be primed by and suffer interference from the presented answer.

Priming from a presented answer in a production task does not directly disconfirm the hypothesis that verification is production plus comparison, and it does not directly support the hypothesis that verification and production tap the same knowledge base in different ways. Instead, such priming bears on the likelihood that subjects will adopt one strategy or the other, and renders the hypotheses more or less plausible in typical verification tasks. In particular, if we find priming, it would make the production-plus-comparison hypothesis less plausible.

Our experiments generalized and extended those of Campbell and his colleagues (Campbell, 1987, 1991; Campbell & Tarling, 1996; Meagher & Campbell, 1995) in two ways. First, we generalized the answer priming effects from multiplication to addition. No one has reported answer priming in addition. Second, we extended the range of SOAs at which priming has been observed to zero, or simultaneous onset, mimicking the conditions typical of verification tasks. Meagher and Campbell found that automatic priming effects were stronger the shorter the SOA, so we might see stronger effects than theirs using a zero SOA.

First we will describe the method, which is common to all the experiments, and then we will present an analysis of variance (ANOVA) that includes data from all the experiments and provides an error term that allows us to compare effects across experiments. Then we will present each experiment individually and build our empirical argument across experiments.

## GENERAL METHOD

### Subjects

Thirty-two subjects served in each experiment. In Experiment 1, 22 subjects were students from an introductory psychology class who participated to fulfill course requirements, and 10 were volunteers from the university community who were paid for their participation. In Experiments 2–4, all the subjects were from the introductory psychology class who participated to fulfill course requirements. Half of the subjects in each experiment (i.e., 16) responded vocally, and half responded manually.

### Apparatus and Stimuli

The stimuli were displayed on Gateway 2000 Crystalscan 1024 NI monitors controlled by Gateway 2000 computers. The subjects responded either by speaking into a voice key attached to the computer's parallel port or by typing on the numeric keypad. Timing was accurate to 1 msec.

There were three displays. The first was blank. The second was a fixation display, which consisted of two rows of five dashes with spaces between them (i.e., - - - - -), one appearing above and one appearing below the line that would contain the upcoming problem. The fixation display was centered on the screen. The third display contained the problem, which consisted of the first single-digit argument, a blank space, a plus symbol (+), a blank space, the second single-digit argument, a blank space, an equal sign (=), a blank space, and a zero-, one-, or two-character answer. In Experiment 1, all of the answers were digits. Half were the true sums of the arguments and half were false sums that were true sums plus or minus 2 (e.g., 3 + 4 = 5; 3 + 4 = 9). In Experiment 2, the true answers were replaced by blanks. In Experiment 3, the true answers were replaced by capital letters. The letters B, C, D, R, G, H, J, K, L, and M were substituted for the digits 0–9, respectively. Thus, when a 0 would have appeared in a true answer, B would appear in its place; when a 1 would have appeared in a true answer, C would appear in its place; and so on. With this arrangement, the statistical redundancy of the letter answers was the same as it would be for the true answers. In Experiment 4, the symbols !, @, #, $, %, ?, &, *, (, and ) were substituted for the digits 0–9, respectively, so that whenever a true answer would contain 0, ! would appear in its place, and so on.

The fixation display was exposed for 500 msec. It was erased and replaced immediately with the problem display, which remained on until the subject responded. In the manual response condition, the subjects typed their answers into the numeric keypad, ending with the "enter" key. The display remained on until the enter key was struck. Then the display was blanked for a 1,500-msec intertrial interval. Manual RT was the interval between stimulus onset and the registration of the first keypress in the series. Logan and Zbrodoff (1998) have shown that Stroop effects with typewritten responses occur only in the latency of the first keystroke and not in the duration of the series of keystrokes. In the vocal response condition, the display was extinguished when the voice key detected a response. The subjects then typed the answer that they had spoken into the numeric keypad, ending with the "enter" key. When they struck the "enter key," a 1,500-msec intertrial interval began.

The fixation display was 1.6 cm high × 2.85 cm wide, and the equations were 5.5 mm × 2.35, 2.85, or 3.1 cm, depending on whether the answer contained zero, one, or two characters. Viewed at a distance of 60 cm, the fixation display was 1.53° × 2.72° of visual angle and the equations were .53° × 2.24°, 2.72°, or 2.96° of visual angle.

### Procedure

Each experiment consisted of 512 trials, formed from two replications of the basic 256-trial design. The 256-trial design was formed by factorially combining 8 possible left addends (digits 2–9), 8 possible right addends (digits 2–9), and 4 answer types. For each combination of digit addends, there were two equations with true answers (e.g., 3 + 4 = 7) and two equations with false answers. There were two types of false answers, true +2 (e.g., 3 + 4 = 9) and true −2 (e.g., 3 + 4 = 5). The false answers were true plus or minus 2 so that their parity would be the same as the true answer (Krueger & Hallford, 1984). Each false problem appeared once with a true +2 answer and once with a true −2 answer. The experiment consisted of two successive replications of the 256-trial design, and the order of trials within replications was randomized separately for each subject.

The subjects were told that their task was to produce the sums of the two digit addends on the left side of the equal sign, ignoring the answer on the right. The vocal response subjects were told to speak loudly and clearly into the voice-key microphone, to type in what they had said on the numeric keypad when the screen went blank after they had spoken, and to press the "enter" key when they had finished. They were told to type in whatever they had said even if they realized it was wrong. The manual response subjects were told to type the sum of the digit addends into the numeric keypad as quickly and accurately as they could and to press "enter" when they had finished. The subjects were allowed brief rests every 64 trials.

### Between-Experiment ANOVA

We submitted the mean RTs and percentage of correct responses to a 4 (experiment) × 2 (answer type: false vs. control) × 2 (re-

sponse type: vocal vs. manual) analysis of variance (ANOVA) on the data from all four experiments. The main effect of experiment was not significant ($F < 1.0$), but the main effect of answer type [$F(1,120) = 120.90$, $MS_e = 811.15$, $p < .01$] and the main effect of response type [$F(1,120) = 16.71$, $MS_e = 70{,}209.10$, $p < .01$] were significant. The only significant interaction was experiment by answer type [$F(3,120) = 6.65$, $MS_e = 811.15$, $p < .01$]. We used the error term from that interaction to construct contrasts that compared answer-type effects between experiments.

## EXPERIMENT 1

### Method

The first experiment mimicked the stimulus conditions in a typical verification task but required subjects to produce correct answers instead of verifying the presented answer. The digit arguments in each problem were presented along with a true or false sum, which appeared in the answer field and was to be ignored. To assess priming, we compared production performance on true answers with production performance on false answers. According to Meagher and Campbell (1995), we could expect a strategic "name-the-prime" effect for true answers and an automatic interference effect for false answers, so the priming effects might be quite large.

### Results and Discussion

The mean RTs to true answer (congruent) and false answer (incongruent) problems were computed for each subject and submitted to a 2 (answer type: true vs. false) × 2 (response type: vocal vs. manual) ANOVA. The means across subjects are presented in Table 1. The percentages of correct responses were computed for each subject and submitted to an ANOVA with the same design. The means across subjects are also presented in Table 1.

Overall, subjects were 56 msec slower to produce the sum of the two digits when they were presented with false answers than when they were presented with true answers. The effect was about the same size for vocal responses (50 msec) and manual responses (61 msec), even though the subjects were 187 msec faster with vocal responses than with manual responses. The difference between true and false answers suggests that subjects were not able to confine the retrieval cues to the digit arguments. The presented answer seems to have had an impact on their ability to recall the correct one. Thus, the data suggest that the format in a typical verification task is likely to discourage a production-plus-comparison strategy and encourage a resonance evaluation strategy that relies on the problem as a whole.

These conclusions were confirmed in the ANOVA done on the mean RTs. The main effect of answer type was significant [$F(1,30) = 39.63$, $MS_e = 1{,}226.98$, $p < .01$], as was the main effect of response type [$F(1,30) = 6.37$, $MS_e = 87{,}312.9$, $p < .05$], but the interaction between answer type and response type was not significant ($F < 1.0$).

The accuracy data were consistent with the RTs. The only significant effect in the accuracy ANOVA was the main effect of answer type [$F(1,30) = 13.48$, $MS_e = 3.51$, $p < .01$].

## EXPERIMENT 2

In Experiment 1, subjects might have paid attention to the equation's answers even though they were instructed to ignore them, because the presented answers were correct on 50% of the trials. The high proportion of correct answers might have encouraged a "name-the-prime" strategy reported by Meagher and Campbell (1995). Experiment 2 was designed to prevent this strategy and to determine whether the difference between true and false answers in Experiment 1 was due to benefits from the true answers, costs from the false answers, or both.

### Method

The design was exactly the same as in Experiment 1, except that true answers were replaced with blanks. Thus, half of the trials involved a verification-like equation with a false answer (e.g., 3 + 4 = 9), and half involved a production-like problem with no answer (e.g., 3 + 4 =). If the answer effect in Experiment 1 was due to strategic attention to the true answers that were presented on half of the trials, there should be no effect of false answers in Experiment 2. When an answer appeared, it was always wrong, so there should be no reason to attend to it strategically. Indeed, Meagher and Campbell (1995) found that removing true answer primes from the stimulus set reduced the magnitude of priming substantially.

### Results and Discussion

The mean RTs to false answer (incongruent) and blank answer (neutral) problems were computed for each subject and submitted to a 2 (answer type: false vs. blank) × 2 (re-

**Table 1**
**Mean Reaction Times (RT) and Accuracy for Incongruent (False Answer), Neutral (Blank, Letter, or Symbol Answer), and Congruent (True Answer) Problems in Experiments 1–4**

| | Experiment 1 (True) | | Experiment 2 (Blank) | | Experiment 3 (Letter) | | Experiment 4 (Symbol) | |
|---|---|---|---|---|---|---|---|---|
| Problem | RT | Acc. | RT | Acc. | RT | Acc. | RT | Acc. |
| | | | Vocal Responses | | | | | |
| Incongruent | 917 | 94.0 | 878 | 94.0 | 951 | 93.9 | 981 | 94.6 |
| Neutral | | | 823 | 94.9 | 932 | 94.9 | 968 | 95.4 |
| Congruent | 867 | 95.8 | | | | | | |
| | | | Manual Responses | | | | | |
| Incongruent | 1,109 | 94.0 | 1,090 | 95.6 | 1,095 | 96.3 | 995 | 95.7 |
| Neutral | | | 1,035 | 96.6 | 1,065 | 97.4 | 963 | 96.9 |
| Congruent | 1,048 | 95.7 | | | | | | |

sponse type: vocal vs. manual) ANOVA. The means across subjects are presented in Table 1. The accuracy data, also presented in Table 1, were analyzed in the same fashion.

Subjects were 55 msec slower to produce their sums when they were presented with false answers than with blanks. The effect was exactly the same size for vocal and manual responses (55 msec), even though subjects were 212 msec faster with vocal responses. The different results for blanks and false answers suggest that subjects were not able to ignore the false answers when retrieving the correct sums. Thus, the data suggest that the presentation of an answer in typical verification tasks discourages the production-plus-comparison strategy and encourages the resonance evaluation procedure.

The RTs were quite similar to the RTs in Experiment 1. The fact that RTs in Experiment 2 were slower for false answers than for blanks suggests that the congruency effect in Experiment 1 was not due entirely to facilitation from the correct answers. The fact that the congruency effect in Experiment 2 (55 msec) was almost identical to the congruency effect in Experiment 1 (56 msec) suggests that there was a cost when false answers were presented in both experiments and no benefit from true answers in Experiment 1.

The within-experiment conclusions were confirmed by the ANOVA of the mean RTs. There were significant main effects of answer type [$F(1,30) = 41.95$, $MS_e = 1{,}151.10$, $p < .01$] and response type [$F(1,30) = 7.80$, $MS_e = 92{,}320.79$, $p < .01$], but the interaction between them was not significant ($F < 1.0$). The accuracy data were consistent with the RTs. The only significant effect was the main effect of answer type [$F(1,30) = 6.50$, $MS_e = 2.16$, $p < .05$].

The between-experiment conclusion was confirmed by a contrast comparing the congruency effects in Experiments 1 and 2, using the error term from the interaction between experiments and answer type. By this criterion, the 56-msec congruency effect in Experiment 1 was not different from the 55-msec effect in Experiment 2 ($F < 1.0$).

## EXPERIMENT 3

The subjects in Experiment 2 might have found it difficult to avoid attending to the false answers because they were mixed with trials in which there was no answer. On many occasions, a trial with a false answer followed a trial with a blank answer, and the false answer might have appeared as a new object, which would have been hard to ignore (see, e.g., Hillstrom & Yantis, 1994). To account for this possibility, in Experiment 3 we replicated the procedure of Experiment 2, presenting capital letters in place of the blank stimuli.

### Method

On half of the trials, false answers were presented to the right of the equal sign, as in the previous experiments (e.g., 3 + 4 = 5), and on the other half, one or two capital letters appeared to the right of the equal sign (e.g., 3 + 4 = H). The idea was to give the subjects something to ignore on every trial, to see whether that would help them to ignore the false answers when they appeared.

Meagher and Campbell (1995) used a similar procedure when they attempted to isolate the automatic priming effect. They removed true answer primes from the procedure, but they presented a prime on each trial. The prime was either a false answer or two symbols (i.e., ##). Thus, Experiment 3 might be more likely than the previous one to assess the true cost of the incorrect answer.

### Results and Discussion

The mean RTs to false answer (incongruent) and letter answer (neutral) problems were computed for each subject and submitted to a 2 (answer type: false vs. letter) × 2 (response type: vocal vs. manual) ANOVA. The means across subjects are presented in Table 1. The accuracy data were analyzed in the same fashion. They are presented in Table 1 as well.

Subjects were 25 msec slower to produce their sums when they were presented false answers than when they were presented letters. The effect was smaller for vocal responses (19 msec) than for manual responses (30 msec), and once again, subjects were faster with vocal responses than with manual responses (by 139 msec). The difference between letters and false answers suggests that subjects were not able to ignore the false answers when retrieving the correct sums.

The congruency effect was smaller in this experiment than in the previous ones (25 vs. 56 and 55 msec), which suggests that subjects found it easier to ignore the answers when half of them were not arithmetically meaningful. Nevertheless, the false answers still had an impact, which suggests that the production-plus-comparison strategy would be difficult to use with the conventional verification-task format.

The within-experiment conclusions were confirmed in the ANOVA on the mean RTs. There were significant main effects of answer type [$F(1,30) = 23.41$, $MS_e = 405.07$, $p < .01$] and response type [$F(1,30) = 5.03$, $MS_e = 60{,}557.16$, $p < .05$], but the interaction between them was not significant [$F(1,30) = 1.21$, $MS_e = 405.07$]. The accuracy data were consistent with the RTs. There were significant main effects of answer type [$F(1,30) = 16.12$, $MS_e = 1.26$, $p < .01$] and response type [$F(1,30) = 6.45$, $MS_e = 14.73$, $p < .05$].

The between-experiment conclusions were tested with a contrast from the between-experiment ANOVA described in the Method section. The contrast between the 25-msec answer type effect in Experiment 3 and the 56- and 55-msec effects in Experiments 1 and 2 was highly significant [$F(1,120) = 48.93$, $MS_e = 811.15$, $p < .01$].

## EXPERIMENT 4

The costs of presenting false answers in the production task were reduced but not eliminated in Experiment 3. It was possible that subjects found it hard to ignore the letters because they were familiar stimuli. To control for this possibility, the letters were replaced by less familiar symbols, and the experiment was replicated.

### Method

On half of the problems, the wrong answer appeared to the right of the equal sign (e.g., 5 + 3 = 10), and on the other half, one or two sym-

bols appeared to the right of the equal sign (e.g., 5 + 3 = @!). Meagher and Campbell (1995) also used symbols as neutral primes, but they presented the same symbols on each neutral trial (i.e., ##).

### Results and Discussion

The mean RTs to false answer (incongruent) and symbol answer (neutral) problems were computed for each subject and submitted to a 2 (answer type: false vs. symbol) × 2 (response type: vocal vs. manual) ANOVA. The means across subjects are presented in Table 1. The accuracy data were analyzed in the same fashion and are also presented in Table 1.

Subjects took 22 msec longer to produce their sums when they were presented with false answers than when they were presented with symbols. The effect was smaller for vocal responses (13 msec) than for manual responses (32 msec). In contrast to the previous experiments, for reasons we do not understand, subjects were only 5 msec faster with vocal responses than with manual responses. More important, the difference between symbols and false answers suggests that subjects were not able to ignore the false answers when retrieving the correct sums.

The congruency effect was about the same in this experiment as it was in Experiment 3 (22 vs. 25 msec), which suggests that subjects did not find it easier to exclude answers when half of them were symbols rather than letters. Thus, symbols may be no less meaningful to arithmetic than letters. The fact that false answers still had an impact suggests that production was not influenced only by the digit arguments, and that makes the production-plus-comparison strategy problematic in standard verification tasks.

The within-experiment conclusions were confirmed in the ANOVA of the mean RTs. There was a significant main effect of answer type [$F(1,30) = 17.05$, $MS_e = 461.95$, $p < .01$], but neither the response type effect ($F < 1.0$) nor the interaction between answer type and response type [$F(1,30) = 3.13$, $MS_e = 461.95$, $p < .10$] was significant. The accuracy data were consistent with the RTs. The only significant ANOVA result was the main effect of answer type [$F(1,30) = 8.82$, $MS_e = 1.81$, $p < .01$].

The between-experiment conclusions were tested with contrasts from the between-experiment ANOVA described in the Method section. The contrast between the 22-msec answer type effect in Experiment 4 and the 56- and 55-msec effects in Experiments 1 and 2 was highly significant [$F(1,120) = 59.03$, $MS_e = 811.15$, $p < .01$]. The contrast between the 22-msec answer type effect in Experiment 4 and the 25-msec effect in Experiment 3 was not significant ($F < 1.0$).

## GENERAL DISCUSSION

In each experiment, subjects were influenced by the presence of false answers, even though they were logically irrelevant to the production task. The subjects in Experiment 1 may have been induced to attend to the answers strategically, because half of the answers were true (Meagher & Campbell, 1995). Experiment 2 ruled out that possibility. True answers were replaced by blanks. Subjects saw only false answers and so should not have been inclined to attend to them strategically. Nevertheless, subjects were influenced by the false answers to the same extent as the subjects were in Experiment 1. Subjects in Experiment 2 might have had their attention captured by the false answers because half of the time the answer field was blank and a false answer appearing in the answer field would often be a new object (Hillstrom & Yantis, 1994). In Experiment 3, we addressed that possibility by presenting letters instead of blanks in the answer field on half of the trials. Subjects were still influenced by the false answers, though to a lesser extent than in Experiments 1 and 2. In Experiment 4, letters were replaced with symbols that were even less arithmetically meaningful, and this resulted in a false answer effect as large as the one in Experiment 3.

Our experiments were intended to force subjects to simulate the production part of the production-plus-comparison strategy in a verification task. We presented them with verification-type problems and asked them to report the sum after they produced it, instead of comparing it with the presented answer and reporting whether or not it matched the correct one. The priming that we observed in each experiment suggests that the answer retrieval process in the production-plus-comparison strategy would suffer from interference from the presented answers. In our view, this interference would induce subjects to look for easier strategies, such as resonance evaluation, which should be facilitated by priming (Campbell, 1991; Zbrodoff & Logan, 1990).

The present results, taken together with those reviewed in the introduction, suggest that resonance evaluation is the prevalent strategy in verification tasks, and the production-plus-comparison strategy is relatively rare. More generally, in suggesting that resonance evaluation prevails in verification, our results and the previous ones support the view that production and verification involve different processes that operate on the same knowledge base (Campbell, 1991; Zbrodoff & Logan, 1990). The results should encourage the development of formal models of the processes and the knowledge base that are analogous to memory models that account for both recognition and recall (e.g., Gillund & Shiffrin, 1984; Murdock, 1993).

Much of the research addressing the relation between production and verification has focused on the underlying processes, proposing a particular stage structure (e.g., production precedes comparison) or a particular retrieval process. The present results suggest that it may be important to focus on differences in stimulus conditions: Verification tasks present true and false answers, whereas production tasks typically do not. The stimulus that is presented may be incorporated in the retrieval cue, whether subjects wish to assess overall activation (as in verification) or to select the most active item (as in production). Future research should address the effects of stimulus conditions in production and verification. In our procedure, the false answers might have been incorporated in the retrieval cue,

because they were grouped together perceptually with the digit arguments. They were similar in form and color, promoting grouping by similarity; they were close to the left-hand side of the equation, promoting grouping by proximity; and they appeared in the answer field, promoting grouping by syntactic structure. Perhaps the effects could be diminished by coloring the answers differently or moving them to more remote, noncanonical locations. Meagher and Campbell (1995) found that the automatic effects could be reduced by separating the arguments and the answer in time.

## CONCLUSIONS

The data from all four experiments show that a presented false answer interferes with production of simple arithmetic sums. The procedure mimicked the production part of the production-plus-comparison strategy for performing verification tasks, and so suggests that the production-plus-comparison strategy would suffer retrieval interference in the standard verification task. The same priming effect would be beneficial for a resonance evaluation strategy because it would add to the difference in activation between true and false items. Thus, the data are consistent with the position that production and verification involve different processes operating on the same arithmetic knowledge base (Zbrodoff & Logan, 1990).

## REFERENCES


Ashcraft, M. H., Fierman, B. A., & Bartolotta, R. (1984). The production and verification tasks in mental addition: An empirical comparison. *Developmental Review,* **4**, 157-170.

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

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

Dagenbach, D., & McCloskey, M. (1992). The organization of arithmetic facts in memory: Evidence from a brain-damaged patient. *Brain & Cognition,* **20**, 345-366.

Geary, D. C., Widaman, K. F., & Little, T. D. (1986). Cognitive addition and multiplication: Evidence for a single memory network. *Memory & Cognition,* **14**, 478-487.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review,* **91**, 1-67.

Groen, G. J., & Parkman, J. M. (1972), A chronometric analysis of simple addition. *Psychological Review,* **79**, 329-343.

Hillstrom, A. P., & Yantis, S. (1994). Visual motion and attentional capture. *Perception & Psychophysics,* **55**, 399-411.

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

Krueger, L. E., & Hallford, E. W. (1984). Why 2 + 2 = 5 looks so wrong: On the odd–even rule in sum verification. *Memory & Cognition,* **12**, 171-180.

Logan, G. D., & Zbrodoff, N. J. (1998). Stroop-type interference: Congruity effects in color naming with typewritten responses. *Journal of Experimental Psychology: Human Perception & Performance,* **24**, 978-992.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103-189). New York: Wiley.

Meagher, P. D., & Campbell, J. I. D. (1995). Effects of prime type and delay on multiplication priming: Evidence for a dual-process model. *Quarterly Journal of Experimental Psychology,* **48A**, 801-821.

Murdock, B. B. (1993). TODAM2: A model for the storage and retrieval of item, associative, and serial-order information. *Psychological Review,* **100**, 183-203.

Parkman, J. M. (1972). Temporal aspects of simple multiplication and comparison. *Journal of Experimental Psychology,* **95**, 437-444.

Parkman, J. M., & Groen, G. J. (1971). Temporal aspects of simple addition and comparison. *Journal of Experimental Psychology,* **89**, 335-342.

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

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. In W. G. Koster (Ed.), *Attention and Performance II* (pp. 276-315). Amsterdam: North-Holland.

Zbrodoff, N. J., & Logan, G. D. (1986). On the autonomy of mental processes: A case study of arithmetic. *Journal of Experimental Psychology: General,* **115**, 118-130.

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, & Cognition,* **16**, 83-97.


(Manuscript received August 17, 1998;
revision accepted February 16, 1999.)