Psychophysiology, 41 (2004), 46–59. Blackwell Publishing Inc. Printed in the USA.

DOI: 10.1111/1469-8986.00119

Comparing arithmetic and semantic fact retrieval: Effects of problem size and sentence constraint on event-related brain potentials

KERSTIN JOST, ERWIN HENNIGHAUSEN, AND FRANK RÖSLER
Department of Experimental and Biological Psychology, Philipps-University Marburg, Marburg, Germany

Abstract

Event-related potentials were recorded with 61 electrodes from 16 students who verified either the correctness of single-digit multiplication problems or the semantic congruency of sentences. Multiplication problems varied in size and sentence fragments in constraint. Both semantic and arithmetic incongruencies evoked a typical N400 with a clear parieto-central maximum. In addition, numerically larger problems (8×7), in comparison to smaller problems (3×2), evoked a negativity starting at about 360 ms whose maximum was located over the right temporal-parietal scalp. These results indicate that the arithmetic incongruency and the problem-size effect are functionally distinct. It is suggested that the arithmetic and the semantic incongruency effects are both functionally related to a context-dependent spread of activation in specialized associative networks, whereas the arithmetic problem-size effect is due to rechecking routines that go beyond basic fact retrieval.

Descriptors: Mental calculation, Problem-size effect, Arithmetic N400 effect, Semantic N400 effect, Memory access, Event-related potentials

Responses to simple mental calculation problems are, in general, slower and more error prone if the operands and hence the correct solutions become numerically larger (e.g., 7×8 compared to 2×3; e.g., Campbell & Graham, 1985; Miller, Perlmutter, & Keating, 1984; Stazyk, Ashcraft, & Hamann, 1982). This so-called problem-size effect holds for both production and verification tasks (Campbell, 1987b; Campbell & Fugelsang, 2001; Parkman, 1972; Zbrodoff & Logan, 1990). Most explanations see its cause in retrieval differences. It is assumed that single-digit problems are solved by direct retrieval of the result from long-term memory. The common theoretical idea is that arithmetic facts are stored in an interrelated associative network linking problems and answers, and it is supposed that the network has similar features as other knowledge networks, for instance, semantic networks of word meaning, and that the theoretical framework of activation spread (Collins & Loftus, 1975) applies here as well (for review, see Ashcraft, 1992, 1995; McCloskey, Harley, & Sokol, 1991). Within this framework, the problem-size effect is attributed to problem-specific differences in the strength with which problems are activated in long-term memory. However, other findings suggest that the prolonged reaction times for problems with large operands could also be due to a substantial use of nonretrieval procedures such as decomposition and reordering (LeFevre, Bisanz, et al., 1996).

In the present study, we will use event-related brain potentials to examine the problem-size effect. In particular, we will study the N400 effect, which is assumed to be an electrophysiological correlate of activation in semantic (Kutas & Federmeier, 2000) and arithmetic (Niedeggen & Rösler, 1999; Niedeggen, Rösler, & Jost, 1999) networks. The arithmetic problem-size manipulation will be compared with a conceptually similar manipulation of the semantic domain, that is, the constraint/cloze-probability variation, to test whether both affect amplitude and topography of the N400 effect in a similar manner.

The Arithmetic Problem-Size Effect

The network model of Ashcraft (1982, 1987) claims that the basic arithmetic facts of addition and multiplication are stored in interrelated memory networks, with problems varying as a function of strength or accessibility. The accessibility of the solutions depends, among others, on the frequency of usage (Ashcraft, 1987). Larger problems are less often encountered in life than smaller problems, and therefore their solutions are less easily activated by the operands than the solutions of smaller problems.

A difference in activation strength for more and less practiced problems is also assumed by Siegler and colleagues. However, in Siegler's Distribution of Associations model (1988; Siegler &

This research was supported by Grant Ro 529/10 from the German Research Foundation (DFG) assigned to Frank Rösler. We thank Patrick Khader for valuable suggestions and helpful discussions.

Address reprint requests to: Kerstin Jost, Fachbereich Psychologie, Philipps Universität Marburg, Gutenbergstraße 18, D-35032 Marburg, Germany; E-mail: jost@staff.uni-marburg.de; or Frank Rösler, Fachbereich Psychologie, Philipps Universität Marburg, Gutenbergstraße 18, D-35032 Marburg, Germany; E-mail: roesler@staff.uni-marburg.de.

Shrager, 1984) calculation problems are not only associated with their correct answer, but also with incorrect solutions that an individual has generated or computed in the past. A problem-size effect results because, in the learning history, more incorrect solutions had been generated for larger than for smaller problems: Faced with an unknown problem, children must resort to strategies other than direct retrieval (e.g., solving 3×4 as $4 + 4 + 4$) and these strategies are assumed to be more error prone for larger than for smaller problems. Given that a particular problem not only activates its correct solution but also incorrect solutions, the distribution of activation will differ for small and large problems. For small problems, the distribution of activation will be more peaked around the correct node than for large problems. The distinctness of a problems activation distribution is assumed to be a key determinant for solution strategies and for behavioral outcomes as percentage of errors and length of solution times.

Another model, proposed by Campbell and colleagues (e.g., Campbell, 1987a, 1995; Campbell & Graham, 1985; Campbell & Oliphant, 1992), stresses the aspect of interference in the network of stored arithmetic facts. A problem not only activates its correct solution, but also a set of other multiples of each operand. Accordingly, incorrect multiples of one of the operands will be produced with higher probability in production tasks (Campbell & Graham, 1985; Miller et al., 1984), and they are harder to reject than other numbers in verification tasks (Campbell, 1987b; Stazyk et al., 1982).

In addition, Campbell (1995) suggests in his modified network interference model that problems also activate a magnitude representation, which reflects the approximate numerical size of the solution. From other work it is known that magnitudes of numbers are related to the objective size by a log linear function, that is, the internal scale is more compressed for larger than for smaller numbers. Thus, larger magnitudes are less well discriminated from each other than smaller magnitudes (see Brysbaert, 1995; Dehaene, 1989; Dehaene, Dupoux, & Mehler, 1990). Therefore, larger solutions are more similar in size to incorrect, table-related solutions. This causes larger problems to activate neighbors more strongly. Consequently, larger problems encounter more interference by way of inhibition received from parallel activated nodes than do smaller problems. This slows down the activation accumulation at the correct node for larger problems, or it may prolong the decision process between correct and incorrect solutions.

The mentioned theories agree on the assumption that differences of activation strength cause the problem-size effect. However, it is not generally accepted that simple arithmetic problems are always solved by fact retrieval. Dehaene and Cohen (1995) distinguish different levels of complexity in calculation tasks, which are assumed to involve different types of subprocesses. They argue that even for single-digit additions and multiplications, normal subjects do not possess a complete and error-free memory (see Campbell & Graham, 1985). Hence, when faced with an unknown or irretrievable fact, subjects may resort to strategies other than retrieval. For instance, a problem may be decomposed into simpler problems for which direct memory representations do exist ($9 \times 8 = (10 \times 8) - (1 \times 8) = 80 - 8 = 72$).

Accordingly, LeFevre and her colleagues (LeFevre, Bisanz, et al., 1996; LeFevre, Sadesky, & Bisanz, 1996) argued that the prolonged reaction times for problems with large operands are due to a substantial use of nonretrieval procedures such as decomposition and reordering.

In the present study, we will try to test by means of event-related potentials (ERP) whether differences in activation spread alone cause the problem-size effect or whether other processes have to be taken into account, too.

ERP Correlates of Activation in Semantic and Arithmetic Memory: The N400 Effects

The N400 effect is a negative deflection in the ERP, which peaks around 400 ms after word onset with a maximum amplitude over centro-parietal brain areas (for review, see Kutas & Van Petten, 1988, 1994). Many studies (Federmeier & Kutas, 1999; Kutas & Hillyard, 1980, 1983, 1984; Kutas, Van Petten, & Besson, 1988) demonstrated that contextually unexpected words in a sentence elicit an N400 effect (“He took his coffee with milk and *mustard*” vs. “… and *sugar*”). Research on the semantic N400 effect has shown that the amplitude depends on the “semantic distance” of a target word from a priming context. For sentence final words the N400 amplitude is largest for completely unrelated words (“He liked lemon and sugar in his *butter*”), intermediate for improbable but related completions (*coffee*) and smallest for the most plausible, best completion (*tea*; Kutas & Hillyard, 1984). Moreover, others have shown that the N400 effect is not restricted to words presented in a sentence context. When presenting lists of isolated words or word pairs in a priming paradigm, the N400 effect proved to be significantly larger for words that had not been preceded by a related word (e.g., Anderson & Holcomb, 1995; Bentin, McCarthy, & Wood, 1985; Chwilla, 1996). Although it is still controversial whether controlled and/or automatic processes contribute to the N400 effect (Brown & Hagoort, 1993; Kiefer & Spitzer, 2000; Rolke, Heil, Streb & Hennighausen, 2001), most findings are compatible with the idea that the amplitude of the N400 effect reflects the amount of additional activation within a semantic network that is necessary to process a just encountered word (Rösler & Hahne, 1992): The more a representation is primed already, the less additional activation is necessary and the smaller the N400 will be.

However, studies with line drawings (Holcomb & McPherson, 1994; Nigam, Hoffman, & Simons, 1992), faces (Barrett & Rugg, 1989), pictures (McPherson & Holcomb, 1999), and sounds (Van Petten & Rheinfelder, 1995) have shown that the N400 effect is not restricted to word stimuli. Incorrect solutions in arithmetic verification tasks also evoke an ERP effect that is comparable to the semantic N400 effect elicited by incongruous words in sentences (Niedeggen et al., 1999). The amplitude of this *arithmetic* N400 effect depends on the relatedness of the incorrect solution with the two operands: The effect is smaller for incorrect solutions, which are associatively connected to one of the operands (incorrect multiples, e.g., $3 \times 5 = 20$) than for completely unrelated solutions (e.g., $3 \times 5 = 19$; Niedeggen & Rösler, 1999). Moreover, the amplitude evoked by operand-related errors depends on the numerical distance between correct and incorrect solutions: It is only attenuated for incorrect results whose magnitude is close to that of the correct product, that is, a direct neighbor. These findings suggest that the amplitude of the arithmetic N400 effect is sensitive to relations between items in long-term memory as is the case for the semantic N400 effect (Kutas & Federmeier, 2000).

The Present Study

Given that the correct solutions of larger problems compared to smaller ones receive weaker activation originating from the operands and are therefore less easily accessible, this should then find an expression in an N400 effect: Correct results of large problems should evoke a negativity comparable to the arithmetic N400 effect elicited by other less or not at all primed stimuli, that is, incorrect results. Consequently, the incongruency N400, that is, the amplitude difference between correct and incorrect solutions, should be smaller for large than for small problems because of the smaller activation difference between correct and incorrect solutions in the case of larger problems. Besides this amplitude difference, no further distinction in the incongruency N400 effects should appear. In particular, N400 topographies should be the same for small and large problems. If, however, the problem-size effect is due to mechanisms other than pure activation differences, for example, that additional strategies such as decomposition and reordering are invoked in solving large problems, then the functional differences between small and large problems should become manifest in a distinct ERP component. These expectations rest on the assumption that ERP effects that differ in amplitude but not in scalp topography result from different strengths of one generator configuration and are therefore indicative of one and the same set of functional processes, whereas qualitative differences as distinct topographies are due to functionally distinct mechanisms (Rugg & Coles, 1995).

Siegler's model (1988) predicts that the use of strategies other than retrieval depends on the relative distribution of associations, that is, these strategies become more important as interference increases. Thus, the problem-size effect could also find an expression in both an amplitude modulation of the arithmetic N400 effect and in an additional ERP phenomenon with a distinct topography.

In the present study we presented correct and incorrect solutions in a verification task (see Niedeggen & Rösler, 1999; Niedeggen et al., 1999). The problem-size effect and its ERP correlates can be studied with this task if the data are analyzed separately for large and for small problems. This results in a design with two factors, problem size with levels small and large, and incongruency of the solution to be verified with levels congruent or correct, and incongruent or incorrect.

The problem-size and the incongruency manipulation in the arithmetic domain can be conceptually related to similar manipulations in the semantic domain. Arithmetic problems establish contextual constraints with respect to the correct solution. According to the activation model outlined above, these constraints should be stronger for small than for large problems. Likewise, a sentence fragment can establish more or less pronounced contextual constraints for a final word. A sentence fragment which triggers, by and large, the very same expectation for the final word in all subjects (e.g., "He takes milk and sugar in his … [*coffee*, *tea*]") is highly constraining. In contrast, a sentence fragment that leaves several options for the final word (e.g., "He takes paper and pencil for his … [*notes*, *drawing*, *sketch*, *exam*, *thesis*, *abstract*, …]") is moderately constraining. Sentences with high and moderate contextual constraints can be determined empirically by collecting cloze probabilities for final words, which are generated as a response to sentence fragments (e.g., Fischler & Bloom, 1979). Words, which are generated with a high or moderate cloze probability (cp), complete a highly or moderately constraining sentence fragment. In addition to the arithmetic conditions, we realized a two-factor design with semantic material, that is, we presented congruent and incongruent endings to sentence fragments that had either a moderate or a high cp. We did so to directly compare the ERP effects of the two materials for the following reason.

For semantic material, it is assumed that sentence constraints and congruency of final words are both functionally related to word expectancy. Thus, for the semantic material, there is no strong argument to assume that the congruency and the constraint manipulation affect distinct mechanisms. In this sense, we expect an equivalent N400 effect for incongruent endings of both highly and moderately constraining sentence fragments and also an equivalent N400 effect for congruent endings, which finalize moderately compared to highly constraining fragments. Because congruent endings are differently preactivated by highly and moderately constraining sentence fragments, the incongruency N400 is expected to differ in amplitude for these two conditions.

To summarize the predictions: (1) Assuming that the arithmetic problem-size effect has its cause in different activation levels of the correct answer nodes, we expect an N400 effect with the same topography for small and large problems but with different amplitudes. Moreover, the topography of this effect should be, by and large, the same as observed for sentence-evoked N400 effects.

(2) If an additional process is invoked by the problem-size manipulation that is functionally distinct from the activation spread mechanism, then topographically distinct ERP effects should become apparent in the difference potentials "large minus small problem size" and "incongruent minus congruent" solutions.

(3) Comparing the arithmetic and the semantic conditions, we expect: The incongruency manipulation of both materials should produce similar N400 effects, whereas the arithmetic problem-size manipulation might produce an additional effect, which is not evoked by the semantic constraint manipulation.

Methods

Participants

Twenty right-handed students of the University of Marburg participated. All were native speakers of German, had normal or corrected-to-normal vision, and were naive with respect to the purpose of the study. They received either course credit points or a monetary compensation. Four participants had to be excluded from the sample because of too many EEG recording artifacts. The final sample comprised data of 10 women and 6 men. The mean age was 24 years (range 21–33 years).

Materials

ERPs in the arithmetic condition were elicited by using single-digit multiplication problems. Ties (e.g., 4×4) were eliminated because they have a processing advantage compared to nontie problems of the same magnitude (Blankenberger, 2001; Campbell & Graham, 1985; Miller et al., 1984). Multiplications involving 0 and 1 were also excluded, as these are assumed to be solved by applying a rule (Sokol, McCloskey, Cohen, & Aliminosa, 1991). Considering these restrictions a set of 56 different operand combinations remained.

Two levels of problem size were defined: Problems with both operands ≤ 5 were classified as small, problems with both operands > 5 were classified as large (see Kiefer & Dehaene, 1997; Zbrodoff & Logan, 1990). Problems with one operand ≤ 5 and one > 5 (e.g., 3×8) were allocated to the two size levels according to the findings of Campbell and Graham (1985) and the classification of Campbell and Tarling (1996), respectively, that is, all problems containing either 2 or 5 were defined as small. Both levels of problem size comprised 28 operand combinations.

For the congruency manipulation, all 56 operand combinations were complemented by the correct and one incorrect solution. Incorrect solutions were either multiples of the first operand or multiplication table unrelated errors. The latter were realized by adding or subtracting 1 from the correct solution.

The orthogonal manipulation of the two factors problem size and congruency results in four experimental conditions: small problem size/correct result (e.g., $3 \times 5 = 15$), small problem size/incorrect result (e.g., $3 \times 5 = 16$), large problem size/correct result (e.g., $9 \times 7 = 63$), and large problem size/incorrect result (e.g., $9 \times 7 = 62$). The whole set of arithmetic problems consisted of 112 trials (2 congruency $\times$ 2 problem size $\times$ 28 operand combinations). To have enough trials for each event category, the total set of trials was repeated once. This resulted in a total of 224 arithmetic trials.

ERPs in the semantic condition were elicited by sentences, each having the syntactic structure of ''noun-verb-direct object'' (e.g., *Karawane durchzieht Wüste* [*caravan passes* (*through*) *desert*]). These triplets had a rudimentary sentence structure, because the noun phrase and the object phrase were both reduced to their heads, that is, they were presented without a determiner that is normally necessary in German to form a fully correct sentence.

Similar to the arithmetic condition, four experimental conditions were constructed by placing both congruent and incongruent endings (two levels of congruency) at the ends of sentence fragments characterized by two levels of contextual constraint. Constraint refers to the extent of expectancy that is built up by the noun–verb phrase; congruency refers to whether the object completed the noun-verb phrase in a meaningful manner. The semantic constraint as an index of context build-up of the sentence fragment was varied (by an operational definition) in terms of cp of the final word (Kutas & Hillyard, 1984). The cloze probabilities were determined in a pilot study with 75 students who were asked in a questionnaire to complete the noun–verb pairs with one other word. Two levels of constraint were defined: Word triplets terminated by a best completion with a cp $> .90$ were defined as highly constraining (average cp $= .96$; e.g., *Gericht fällt* Urteil [*law-court passes* sentence]), those terminated by a best completion with a cp $\leq .90$ were defined as moderately constraining (average cp $= .81$; e.g., *Ritter erobert* Burg [*knight conquers* castle]). Both constraint levels comprised 28 noun–verb pairs and each pair was combined with both a congruent and an incongruent completion. For congruent word triplets we always used the word that had been generated with the highest probability (e.g., Ritter erobert *Burg* [knight conquers *castle*]). Incongruent word triplets were constructed by completing the noun–verb pair with a word that had never been generated in the cloze procedure (by definition cp $= 0$; e.g., Ritter erobert *Geräusch* [knight conquers *noise*]). The average frequency of the terminal words (determined from the CELEX database; Centre for Lexical Information, 1995) was equalized for each of the four conditions. An analysis of variance (ANOVA) with the word frequencies as a dependent variable revealed no significant main effect of condition. As in the arithmetic condition, the total set of word triplets (2 congruency $\times$ 2 constraint $\times$ 28 noun–verb pairs $= 112$) was repeated once. This resulted in 224 semantic trials.

To assess the concreteness of our stimulus material, 23 participants rated the concreteness of all target words on a 5-point scale. The average scores amounted to 4.14, 3.70, 4.00, and 3.64 for targets in the high constraint/congruent, high constraint/incongruent, moderate constraint/congruent, and moderate constraint/incongruent condition. An ANOVA revealed that terminal words in highly constraining sentences were evaluated on average as more concrete than terminal words in moderately constraining sentences, $F(1,22) = 10.42$, $p = .0039$, and words used for the congruent sentences were rated as more concrete than those used for incongruent sentences, $F(1,22) = 131.10$, $p < .0001$.

Procedure

Participants sat in an electrically shielded, dimly lit, and sound-attenuating room in front of a computer screen (ATARI SM 124; refresh rate, 70 Hz) located at eye level at a distance of 70 cm. Alphanumeric characters were presented in black on a light-gray background and subtended a visual angle of 0.4° (vertical size, 0.5 cm).

At the beginning of each trial, a fixation cross was presented in the center of the screen for 250 ms. Participants were instructed to suppress eye movements and blinks from this point on. After an interval of 500 ms, the first word or operand was presented. A trial consisted of three words or three numerals, which were presented successively and separated by an ISI of 250 ms. Multiplication problems were presented without a multiplication sign or equation mark. The two operands and the first two words, respectively, were visible for 350 ms. The presentation of the solution and the terminal word was terminated by the response. The intertrial interval varied between 2 and 3 s.

Participants were instructed to decide on the correctness of the multiplication problems or the congruency of a sentence. They were instructed to respond as quickly and accurately as possible.

The response was given by a brief upward finger movement, which interrupted a light gate. Allocations of the two index fingers to the two response alternatives were varied systematically across participants. Participants had to decide within an interval of 1,500 ms after onset of the critical stimulus. Delayed and incorrect responses were fed back. These trials were excluded from the analysis.

The arithmetic and the semantic materials were presented blockwise. Half of the participants began with the semantic condition, the other half with the arithmetic condition. The total set of 448 trials was split into 16 blocks with 28 trials each. The trials within each condition were presented in random order with the restriction that neither the same multiplication problem nor the same noun–verb pair was repeated within a block. Congruent and incongruent stimuli appeared equally often within a block. A new trial sequence was created for each participant. An obligatory break of at least 20 s separated the blocks. Prior to the experimental trials, participants received one block of arithmetic problems and one block of sentences, each containing 12 trials. Sentences and multiplication problems in these practice blocks were not used for the experiment proper. In total, the experiment took about 1 hr.

EEG Recording, Artifact Handling, and Signal Extraction

The EEG was recorded with 61 Ag/AgCl electrodes inserted in an elastic cap (easy cap, Falk Minow Systems, Munich, Germany) with predefined electrode positions, extrapolated from the 10–20 system (Jasper, 1958). The electrodes were referenced to the nose tip. Additional electrodes were attached at the outer canthi of both eyes and the sub- and supraorbital ridges of the left eye for horizontal and vertical EOG recording. The left mastoid served as the ground. Electrode positions were abraded with OmniPrep and filled with electrolyte gel. Impedances were always kept below 5 kΩ.

EEG and EOG were recorded using two sets of 32-channel amplifiers (SYNAMPS, NeuroScan). Data acquisition, stimulus presentation, and response recording were computer controlled. Band pass of the recording system was set from DC to 30 Hz with a digitization rate of 200 Hz. A DC reset was initiated automatically prior to the beginning of each experimental block.

Trials were rejected if the average amplitude exceeded 125 μV in one of the channels. Trials containing eyeblinks were detected by a wavelet analysis and rejected. Drift artifacts were corrected according to a method suggested by Hennighausen, Heil, and Rösler (1993). From the edited set of raw data, ERPs were extracted for each subject by averaging single trials separately for electrodes and each experimental condition. For this, only trials with correct responses were used. Taking together all mentioned corrections, the percentage of rejected trials never exceeded 20% in one subject and experimental condition. Therefore, averages are always based on a minimum of 45 trials.

Dependent Variables and Statistical Analysis

Behavioral data. Mean response times and error rates were computed for each participant and experimental condition. Trials with response times shorter than 200 ms were excluded. By this procedure less than 1% of the trials were eliminated.

The data were submitted to an ANOVA with factors type of material (arithmetic, semantic), congruency (congruent, incongruent), and constraint (small vs. large problem size and high vs. moderate constraint).

EEG data. For the statistical analysis of the EEG data, we selected 12 standard electrodes to avoid an inflation of the degrees of freedom in the analysis with factor electrode. This selected set of electrodes—F7, F3, F4, F8, T3, C3, C4, T4, T5, P3, P4, T6—allows the scalp sites to be split into three orthogonal factors of caudality (anterior, central, posterior), laterality (left, right), and verticality (medial, lateral). A topographic analysis with these three spatial factors was computed only if the interaction with the unsorted factor electrode had revealed a significant effect.

Average amplitudes were computed for 26 consecutive time windows of 30-ms length (six sampling points) starting with the onset of the solution of the arithmetic problem or the final word of the word triplet. Such narrow time windows increase the sensitivity to detect differences between conditions and topographies. The average amplitudes were measured relative to a baseline of 100 ms, which preceded the onset of the critical stimulus.

A first inspection of the data showed that the time course of the incongruency and constraint effects was different for the two types of material. Therefore, we defined separate sets of ANOVAs for the arithmetic and the semantic material to test the influence of the experimental manipulations. In a first step, superordinate ANOVAs were run for each time window considering the factors congruency, constraint, and electrode. After this, "local" ANOVAs were computed for each time window and electrode location given that the superordinate ANOVA had signaled a significant interaction with factor electrode (see Rösler, Friederici, Pütz, & Hahne, 1993, for the hierarchical testing rationale). Amplitude differences were considered as reliable only if at least two consecutive time windows had signaled significance ($p < .05$) for the particular effect. Marginal effects or effects of short duration (extending over one time window of 30-ms length only) are reported only if they are relevant for the a priori hypotheses. All ANOVA factors were defined as repeated measures. F statistics were corrected according to the formulas of Greenhouse and Geisser (1959). The uncorrected degrees of freedom, the corrected p value, and the respective epsilon values are reported.

To compare topographies, we first computed average difference amplitudes of the time window in which the maximum effect had been observed. These data were standardized such that the mean and variance across electrodes in each condition is zero and one, respectively. The standardization permits evaluating topographic differences between conditions that are not caused by amplitude or variance differences (see McCarthy & Wood, 1985; Urbach & Kutas, 2002).

Results

Behavioral Data

The average response times and error rates of both materials are summarized in Table 1. The ANOVA with factors material (arithmetic, semantic), constraint (small vs. large problems or high vs. moderate constraint sentences), and congruency of target (correct vs. incorrect solution or congruent vs. incongruent final word) confirm that the experimental manipulations had been successful: Responses for incongruent trials were slower than for congruent trials (main effect congruency, $F(1,15) = 30.68$, $p < .0001$). This difference holds for both materials ($p < .05$) but is larger for the semantic (62 ms) than

Table 1. *Average Response Times and Error Rates (±Standard Deviation) of the Four Conditions of the Arithmetic and the Semantic Verification Task*

Material	Constraint of context	Congruency of target	Response time (ms)	% Error
Arithmetic (multiplication problems, the target was a correct or an incorrect solution)	Small problem size	Correct	517 (±53)	3.59 (±4.22)
		Incorrect	560 (±47)	2.79 (±2.84)
	Large problem size	Correct	606 (±78)	9.95 (±7.45)
		Incorrect	627 (±66)	7.43 (±7.46)
Semantic (three-word sentences, the target was a congruent or an incongruent final word)	High constraint	Congruent	511 (±55)	1.11 (±2.04)
		Incongruent	581 (±44)	2.23 (±2.00)
	Moderate constraint	Congruent	544 (±58)	1.33 (±1.66)
		Incongruent	599 (±46)	1.34 (±1.79)

for the arithmetic task (33 ms, Material × Congruency, $F(1,15) = 9.80$, $p = .0069$). The problem size/constraint manipulation affected response times, too (main effect constraint, $F(1,15) = 73.10$, $p < .0001$). Further analyses showed that problem size in the arithmetic task had a more pronounced effect on response time than constraint in the semantic task (Material × Constraint, $F(1,15) = 17.42$, $p = .0008$). The overall response time difference between small and large problems amounted to 78 ms whereas the equivalent difference between moderately and highly constraining sentences was only 26 ms (both $p < .05$). Bonferroni corrected pairwise comparisons within each material showed that all average response times (RTs) differ significantly from each other with one exception: In the arithmetic condition, RTs to correct and incorrect solutions of large problems are statistically equivalent (difference between 606 and 627 n.s.).

The corresponding analysis of error rates revealed that participants on average committed less errors in the semantic than in the arithmetic condition (main effect material, $F(1,15) = 19.59$, $p = .0005$). Furthermore, error rates were affected only by the problem-size manipulation in the arithmetic condition, not by the constraint manipulation in the semantic condition (Material × Constraint: $F(1,15) = 12.84$, $p = .0027$). Subjects committed more errors with larger than with smaller problems ($p < .05$). Other than for response times, factor congruency had no reliable effects on error rates.

Event-Related Potentials

For the statistical analyses, the average amplitude was measured for subsequent time windows of 30-ms length, starting with target onset, that is, the presentation of the solution of the arithmetic problem and the final word of the word triplet. To account for latency differences of the effects, ANOVAs were run for each time window separately for the arithmetic and the semantic conditions. These ANOVAs considered the factors constraint of priming context, congruency of target, and electrode.

Arithmetic Conditions

The grand averages of the original ERPs, which were evoked by correct and incorrect solutions of small and large problems, are presented in Figure 1 (upper part). Figure 2 (upper part) visualizes the incongruency and the problem-size effects, each computed as the difference of two respective grand average ERPs.

Incorrect solutions compared to correct solutions evoked a relative negativity starting at about 270 ms that reached its maximum effect size of approximately $-7\,\mu$V at about 350 ms over the centro-parietal scalp (Figure 1). The maximum is reached later with larger than with smaller problems whereas the amplitude is of the same size in both conditions (Figure 2, upper part). The relative negativity resolves completely between 450 and 500 ms and after this a relative positivity prevails until about 700 ms. The bipolar waveform holds for small and large problems and is comparable to the waveform described by Niedeggen et al. (1999). Therefore, we address the relative negativity as arithmetic N400 effect, or more precisely in the present context, as arithmetic incongruency N400 effect.

Problem size affected the ERPs at a much later point in time than incongruency (Figures 1 and 2). Larger problems evoked a relatively more negative potential than smaller problems for likewise correct and incorrect solutions. This difference starts at about 350 ms and prevails until the end of the recording epoch (800 ms). The effect seems to be more pronounced over the right than over the left hemisphere (Figure 2). Statistical analyses will reveal that the timing and the topography of this problem-size effect is not equivalent to the typical arithmetic N400 effect as was found for the incongruency manipulation. These visual impressions are substantiated by the statistics.

Incongruency effect. Superordinate ANOVAs revealed a main effect congruency for all time windows between 270 and 420 ms, min$F(1,15) = 7.93$, $p = .0131$; max$F(1,15) = 62.87$, $p = .0001$, and between 450 and 660 ms, min$F(1,15) = 5.57$, $p = .0322$; max$F(1, 15) = 29.11$, $p = .0001$. These two intervals cover the negativity and the following positivity in the difference waves shown in Figure 2 (upper part). The slight latency shift between small and large problems is reflected by the interaction Problem Size × Congruency that is marginally significant between 270 and 330 ms, min$F(1,15) = 3.14$, $p = .0967$; max$F(1,15) = 4.05$, $p = .0625$, and again between 390 and 420 ms, $F(1,15) = 3.35$, $p = .0872$. For the effect maximum, between 330 and 360 ms, neither the interaction Problem Size × Congruency nor the interaction Problem Size × Congruency × Electrode was significant. Thus, problem size affected the latency but neither the amplitude nor the topography of the incongruency effect.

The interaction Congruency × Electrode between 270 and 450 ms, min$F(11,165) = 5.33$, $p = .0024$, $\varepsilon = .2921$; max$F(11,165) = 15.67$, $p = .0001$, $\varepsilon = .3328$, justifies a detailed analysis of the topography of the incongruency N400 effect. Electrode-wise analyses revealed a significant congruency effect ($p < .05$) at all electrodes between 300 and 390 ms with a maximum effect size predominated at medial centro-parietal electrodes (C3, C4, P3, and P4; see Figure 2). A more detailed topographic analysis with three spatial factors caudality (anterior, central, posterior), laterality (left, right), and verticality (medial, lateral) confirmed that the incongruency effect is larger at central and posterior than at anterior scalp sites (Congruency × Caudality, min$F(2,30) = 11.68$, $p = .0015$, $\varepsilon = .6463$; max$F(2,30) = 22.11$, $p = .0001$, $\varepsilon = .6650$) and also larger at medial than at lateral sites (Congruency × Verticality, min$F(1,15) = 25.48$, $p = .0001$; max$F(1,15) = 65.87$, $p = .0001$) for the time windows between 300 and 390 ms.

The topographic analysis of the following positivity revealed a significant interaction Congruency × Caudality × Verticality between 510 and 600 ms, min$F(2,30) = 4.39$, $p = .0233$, $\varepsilon = .9490$; max$F(2,30) = 7.81$, $p = .0047$, $\varepsilon = .7632$, showing that the effect was most pronounced at medial posterior electrodes (P3, P4; see Figure 2). This is the same topography as was found by Niedeggen and Rösler (1999).

Problem-size effect. Larger problems evoked a relatively more negative potential than smaller problems between 360 and 780 ms (main effect problem size: min$F(1,15) = 7.38$, $p = .0159$; max$F(1,15) = 19.87$, $p = .0005$). This effect does not interact with the incongruency manipulation after 420 ms. There is only a slight difference in onset latency (see Figure 2), that is, the difference between small and large problems starts to become significant at 360 ms for incorrect solutions and at 390 ms for correct solutions (all $p < .05$).

With respect to the topography, the interaction Problem Size × Electrode proved significant between 420 and 510, min$F(11,165) = 4.16$, $p = .0112$, $\varepsilon = .2711$; max$F(11,165) = 5.21$,

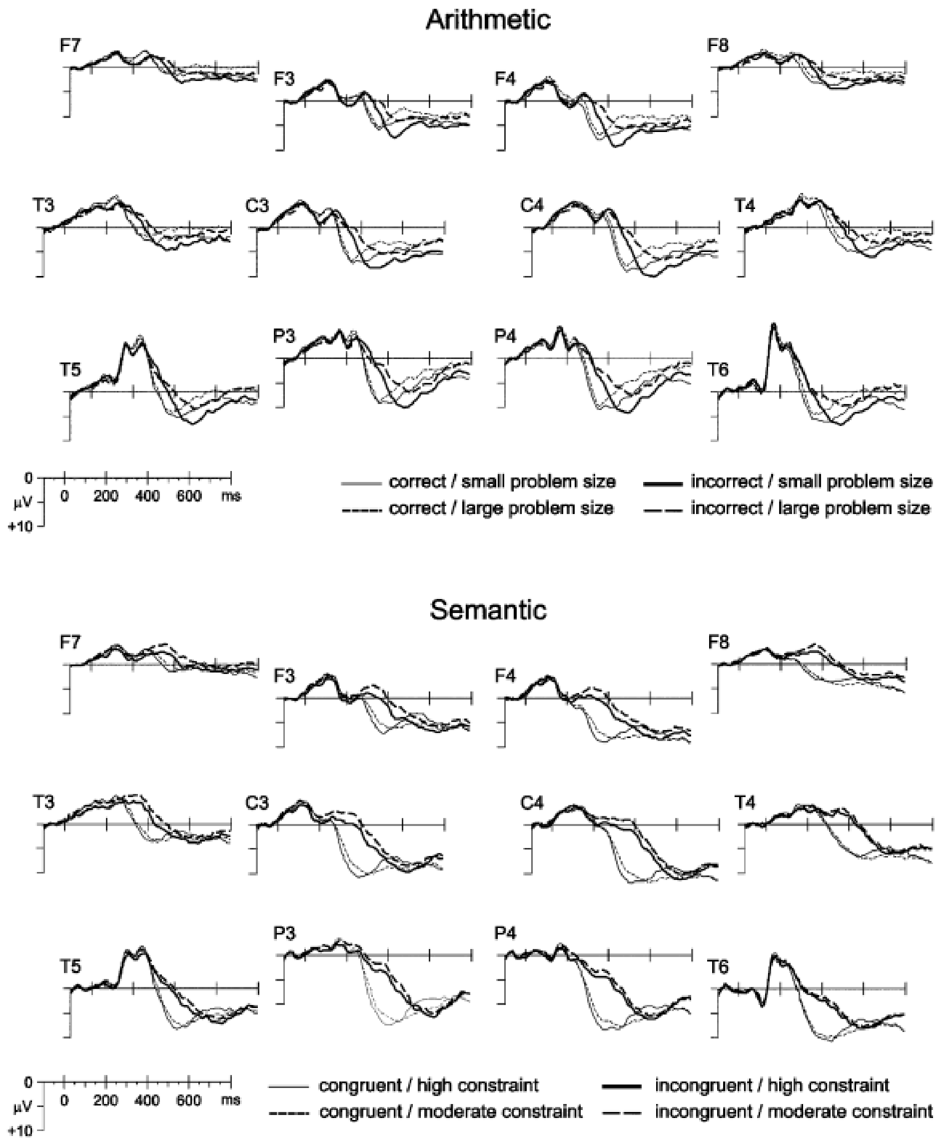

Figure 1. Grand average ERPs for the 12 analyzed electrodes. Upper part: ERPs evoked by the four arithmetic conditions, that is, correct and incorrect solutions presented contingent to small and large problems. Lower part: ERPs evoked by the four semantic conditions, that is, congruent and incongruent final words presented contingent to highly and moderately constraining sentence fragments. In this and the following figures all ERPs are referenced to a 100-ms-long baseline preceding the onset of the critical stimulus (0 ms) and negativity is up.

$p = .0035$, $\varepsilon = .2870$). The effect could be confirmed with post hoc tests for all electrodes (all $p < .05$), but a closer look reveals that the effect is most pronounced at C4, T4, P4, and T6. This clear lateralization is also confirmed by the more detailed topographic analysis for the epoch between 420 and 510 ms (Problem Size × Laterality, $\min F(1,15) = 5.55$, $p = .0325$; $\max F(1,15) = 18.69$, $p = .0006$).

A direct comparison of the topographies of the maxima of the incongruency effect (330–360 ms) and of the problem-size effect (450–480 ms) substantiates the striking topographic difference. Figure 2 (right side) shows the standardized maps. They disclose that the problem-size effect has a clear lateralization to the right in comparison to the incongruency effect. This difference is confirmed by the analyses with unstandardized (Type of Effect × Electrode, $F(11,165) = 4.80$, $p = .0037$, $\varepsilon = .3079$) and with standardized z scores (Type of Effect × Electrode, $F(11,165) = 4.26$, $p = .0044$, $\varepsilon = .3596$).

In summary, the congruency and the problem-size manipulation became manifest in the ERPs. Incongruency furnished a typical arithmetic N400 effect around 350 ms with a central-parietal maximum. Problem size affected the peak latency but not the amplitude and the topography of the incongruency N400. Further, large problems compared to small problems evoked a relative negativity for correct and incorrect equations. This problem-size effect peaked later in time than the incongruency effect and had a clear lateralization to the right hemisphere. Due to differences in timing and topography, incongruency and problem-size effect seem to be functionally distinct.

Semantic Conditions

Figure 1 (lower part) shows the ERPs, which are evoked by congruent and incongruent final words of highly and moderately constraining sentence fragments. Figure 2 (lower part) visualizes the incongruency effect (incongruent minus congruent) sepa-

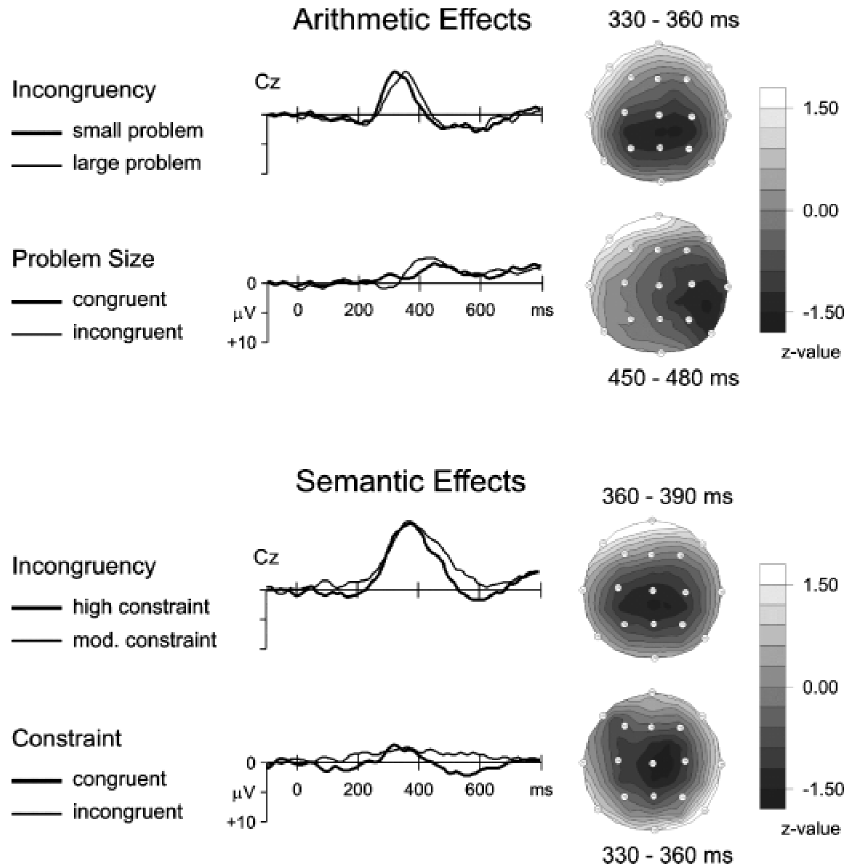

Figure 2. Difference waves at the left side showing the time course of the arithmetic (upper part) and the semantic (lower part) effects for one electrode (Cz). The top row shows the arithmetic incongruency N400 effect separately for small and large problems. Difference waves were calculated by subtracting point by point the ERP of correct solutions from that of incorrect solutions. Below are shown the problem-size effects (large minus small) separately for correct and incorrect solutions. The difference waves in the lower part show the respective semantic effects, that is, the incongruency N400 effect (incongruent minus congruent) separately for high and moderate constraint sentences and the constraint effects (moderate constraint minus high constraint) separately for congruent and incongruent sentences. The *z* maps on the right side show the topographic distribution of the respective effects for the time windows of maximum effect size. Maps are computed from the average amplitudes of the indicated 30-ms-long epoch using a linear interpolation. Darker shading indicates a more pronounced relative negativity.

rately for the high and moderate constraint condition and the constraint effect for congruent and incongruent sentences. Incongruent in comparison to congruent final words evoked a widely distributed relative negativity starting at about 250 ms, which reached its maximum size of approximately $-10\,\mu$V at about 380 ms over the centro-parietal scalp (Figure 1). Rising flank and maximum are identical in latency and amplitude for both constraint conditions (Figure 2). The relative negativity resolves earlier for high constraint (around 510 ms) than for moderate constraint sentences (around 600 ms). In the high constraint condition, a positivity follows the negativity that prevails between 500 and 700 ms. The relative negativity has the timing and the topography of a typical semantic N400 effect.

Sentence constraint (Figure 1) affected the ERPs only little, but, nevertheless, final words in moderately constraining sentences evoked a relative negativity in comparison to final words in highly constraining sentences (approximately $-2\,\mu$V). This relative negativity follows both congruent and incongruent words, but extends longer in time (until 600 ms) for the latter than for the former (until 450 ms, see Figure 2).

ANOVAs were again run with average amplitudes of subsequent time windows of 30-ms width starting with target onset.

Incongruency effect. The superordinate ANOVAs revealed that the incongruency N400 effect is reliable between 270 and 480 ms, $\min F(1,15) = 16.36$, $p = .0011$; $\max F(1,15) = 105.09$, $p = .0001$. The incongruency effect is statistically equivalent for the high and moderate constraint condition in this interval. Differences between highly and moderately constraining sentences are significant in the resolving flank between 480 and 690 ms (Constraint × Congruency, $\min F(1,15) = 4.85$, $p = .0437$; $\max F(1,15) = 13.00$, $p = .0026$). In the moderate constraint condition, the incongruency effect prevailed significant as negativity until 540 ms, $\min F(1,15) = 9.42$, $p = .0078$; $\max F(1,15) = 23.59$, $p = .0002$, whereas in the high constraint condition, the N400 effect was reliable only until 480 ms, followed by a positivity that proved significant between 570 and 630 ms, $\min F(1,15) = 5.53$, $p = .0328$; $\max F(1,15) = 7.50$, $p = .0152$.

A nonuniform distribution of the incongruency effect over the scalp is signaled by the interaction Congruency × Electrode between 270 and 540 ms, $\min F(11,165) = 3.22$, $p = .0200$, $\varepsilon = .3493$; $\max F(11,165) = 14.89$, $p = .0001$, $\varepsilon = .2836$. The incongruency N400 has a very broad distribution, that is, the effect is reliable ($p < .05$) at all electrodes between 300 and 510 ms, but a clear topographic maximum is located over centro-parietal electrodes (C3, C4, P3, and P4), which is also confirmed by the detailed topographic analysis (Congruency × Caudality × Verticality, min$F(2,30) = 3.93$, $p = .0457$, $\varepsilon = .7396$; max$F(2,30) = 8.81$, $p = .0010$, $\varepsilon = .9597$).

The positivity that can be seen in the difference wave of the high constraint condition between 600 and 630 ms (see Figure 3) has an effect maximum over the posterior part of the scalp (Congruency × Caudality, $F(2,30) = 5.99$, $p = .0199$, $\varepsilon = .6085$).

Constraint effect. The main effect of constraint is reliable between 300 and 450 ms, min$F(1,15) = 5.53$, $p = .0328$; max$F(1,15) = 15.53$, $p = .0013$, and, as already mentioned, the interaction Congruency × Constraint is significant between 480 and 690 ms. Separate analyses for congruent and incongruent final words show that the constraint effect is significant for incongruent words between 300 and 600 ms, min$F(1,15) = 5.74$, $p = .0301$; max$F(1,15) = 10.98$, $p = .0047$. For congruent words, the effect remains negative until 480 ms and reverses later on into a small positivity that differs reliably from zero between 540 and 600 ms, min$F(1,15) = 6.65$, $p = .0210$; max$F(1,15) = 6.03$, $p = .0267$.

With respect to the topography of the constraint effect, the interaction Constraint × Electrode proved significant between 300 and 390 ms, min$F(11,165) = 3.53$, $p = .0259$, $\varepsilon = .2501$; max$F(11,165) = 8.20$, $p = .0001$, $\varepsilon = .3055$. The effect is significant with $p < .05$ at all electrodes except T3, T4, T5, and T6, which is also confirmed by the detailed topographic analysis (Constraint × Verticality, min$F(1,15) = 23.03$, $p = .0002$; max$F(1,15) = 32.75$, $p = .0001$). This pattern does not differ for congruent and incongruent sentences.

For a direct comparison of the incongruency and the constraint effect, we selected the amplitudes of the maximum effect size (i.e., 360–390 ms for the incongruency and 330–360 ms for the constraint effect). The ANOVA of the original amplitudes indicated a significant difference (Type of Effect × Electrode, $F(11,165) = 9.50$, $p = .0002$, $\varepsilon = .2242$. The incongruency N400 effect had a larger amplitude than the constraint effect at all electrodes, and these amplitude differences are most pronounced over medial sites of the central-parietal cortex (Type of Effect × Verticality: $F(1,15) = 20.41$, $p = .0004$; Type of Effect × Caudality: $F(2,30) = 13.31$, $p = .0014$, $\varepsilon = .5778$). The standardized topographies of the effects are shown in Figure 2 (right part). As can be seen, the topographic maximum of both the incongruency and the constraint effect is located over the central part of the scalp and the overall topography is very similar. Nevertheless, the ANOVAs computed with standardized z scores indicated a topographic difference (Type of Effect × Electrode: $F(11,165) = 5.51$, $p = .0007$, $\varepsilon = .3748$), which is probably due to the different extension of the negative peaks.

In summary, incongruent final words evoked a typical N400 effect between 300 and 500 ms with a centro-parietal maximum. Sentence constraint had an additional effect on the ERPs within the same time epoch, but with smaller overall amplitude. Despite some differences in their spatial distribution, both effects showed a clear topographic maximum over the central parietal scalp. Both manipulations did not interact during the epochs of their maximum effect size. An interaction was found with respect to the duration of the incongruency N400, which was prolonged in the condition with moderate constraint. Moreover, it was only in the high constraint condition that the N400 effect was followed by a positive deflection.

Topographic Differences between Arithmetic and Semantic Conditions

The striking similarity of the arithmetic and the semantic incongruency N400 effects becomes obvious in Figure 3. Onset latency is almost the same for both materials, whereas the maximum amplitude is, in general, smaller and reached earlier in the arithmetic than in the semantic condition. Nevertheless, the topographies in the time windows of maximum effect size are similar (see Figure 2). As indicated by the analyses, both show a wide distribution with a central-parietal maximum. A direct comparison of the maximum amplitudes run with the original difference scores revealed significance (Material × Electrode, $F(11,165) = 3.26$, $p = .0355$, $\varepsilon = .2448$). However, the ANOVA run with z scores gave no evidence for topographic differences.

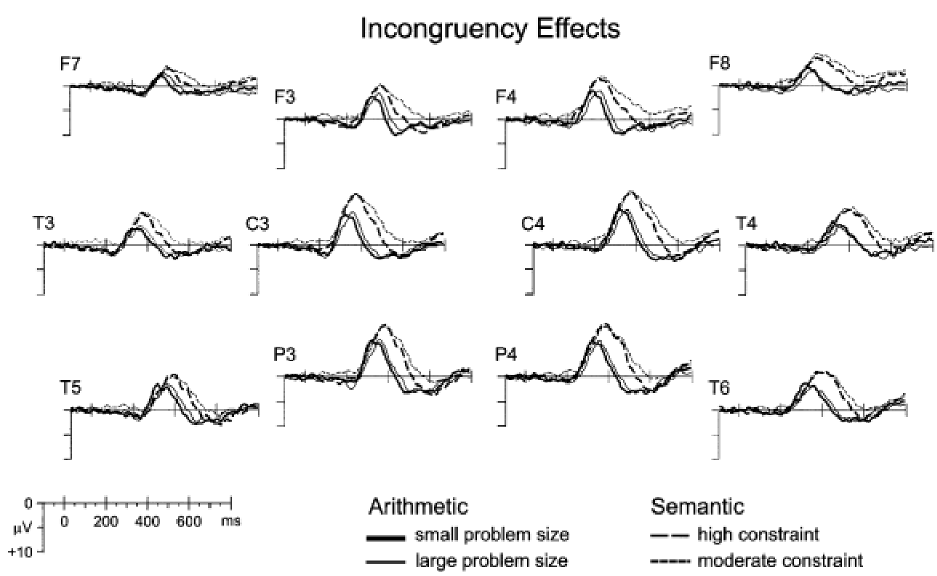

Figure 3. Arithmetic and semantic incongruency effects (incongruent minus congruent) separately for the two levels of problem size/sentence constraint for the 12 analyzed electrodes.

Thus, the difference seen with the original amplitudes is most likely due to an amplitude modulation of one and the same set of generators (McCarthy & Wood, 1985; Urbach & Kutas, 2002).

The positivity following the incongruency N400 was reliable for small and large problems and for highly, but not for moderately constraining sentences. Despite this difference, the topography of the positivity was almost the same in the semantic and the arithmetic condition. Neither the analysis computed with the original amplitude differences nor the analysis computed with z scores revealed a significant effect.

For completeness, the topographies of the semantic constraint effect and of the arithmetic problem-size effect were also compared. As expected, the ANOVAs run with unstandardized (Material $\times$ Electrode, $F(11,165) = 4.57$, $p = .0048$, $\varepsilon = .3103$) and with standardized z scores (Material $\times$ Electrode: $F(11,165) = 5.22$, $p = .0012$, $\varepsilon = .3612$) revealed significant topographic differences.

Discussion

Verification of Multiplication Problems

The objective of the present study was to test whether the arithmetic problem-size effect is functionally related to spread-of-activation differences in a memory network, which stores arithmetic facts. This hypothesis was contrasted with the alternative that small and large one-digit multiplication problems invoke distinct solution strategies.

Assuming that a presentation of a multiplication problem is functionally equivalent to a priming condition, in that the operands prime the solution in an associative network as words or concepts prime other words and concepts in a semantic memory network, one can predict that solutions of larger problems should receive less activation from the priming operands than solutions of smaller problems. This difference in activation spread should find an expression in the N400 amplitude. In contrast, if the problem-size effect has another functional basis than activation spread alone, for example, different solution strategies, then this should become apparent in an ERP effect different to the typical N400 effect.

The observed pattern of results clearly suggests that the arithmetic problem-size effect and the arithmetic incongruency effect cannot be attributed to the same functional mechanism. Both manipulations prolong decision times comparable to the findings in other studies (e.g., Campbell, 1987b; Campbell & Fugelsang, 2001; Zbrodoff & Logan, 1990), but they become manifest in ERP components with distinct topographies.

The incongruency effect was accompanied by a negativity around 350 ms with a topography similar to those found for incongruent sentences (see below). In contrast, correct solutions of large problems compared to correct solutions of small problems evoked an additional negativity, which peaked later and with a different topography than the typical arithmetic N400 effect. The maximum was found around 450 ms over the right central to parietal cortex. This effect was also observed for incongruent results.

An amplitude modulation by problem size was found not only in late time windows, but also for the time epoch of the N400 effect. However, problem size did not affect the amplitude, as originally expected, but the latency of the N400 effect: The peak amplitude was reached later for larger than for smaller problems, although the topography was the same. One possible explanation for this unexpected latency shift could be that it results from an overlap with a P300 that shifts its latency between the conditions. As is obvious from Figure 1, the negativity elicited by incorrect results is followed by a positive wave, which seems to reach its maximum amplitude later in time for larger than for smaller problems. However, the latency difference between the incongruency N400 of small and large problems was not only present in the resolving flank, but also in the rising flank. This makes it unlikely that it was caused by an overlapping positivity with different onset or peak latencies. This raises the question about the cause of the latency shift.

Our predictions took into account differences in the strength but not in the temporal dynamics of activation spread in arithmetic memory networks. The latter aspect is outlined in Campbell's modified network-interference theory (for a similar assumption concerning verification tasks, see the resonance theory by Zbrodoff & Logan, 1990; Campbell & Tarling, 1996). There it is assumed that not only the amount of interference is greater for larger problems, leading to a less peaked distribution, but also that the time needed to produce a response should depend on the distribution of activation. The correct node for a problem with larger operands receives a relatively smaller proportion of the total activation, and this should result in a small initial signal-to-noise ratio between correct and incorrect solutions. Given, as in Campbell's simulation, that the signal-to-noise ratio increases with processing time, correct nodes of large problems will need more time to accumulate enough activation to reach a specified threshold. This could be a basis of the observed N400 latency differences between small and large problems.

The same reasoning should also apply to the semantic domain. According to Schwanenflugel and Shoben (1985), highly constraining sentence fragments increase the expectation for certain words while simultaneously decreasing the expectation for other words, that is, high constraint sentences should build up a more peaked activation distribution than moderate constraint sentences. Accordingly, this should also find an expression in the time course of the N400. However, to our knowledge, it has not yet been observed that the N400 shifts its peak latency as a function of sentence constraint. On the other hand, in the present study, the semantic N400 effect showed a timing difference too: The incongruency N400 resolved earlier for highly constraining sentences than for moderately constraining sentences, which could be a consequence of a different activation ratio between fitting and nonfitting sentence endings. This post hoc explanation leaves open, however, why the timing difference appeared in the rising flank with arithmetic material and in the resolving flank with sentence material. At present, we have no convincing argument to explain this discrepancy.

Irrespective of the unexpected timing differences, the most interesting finding is the late problem size effect with a maximum over the right cortex. As outlined in the introduction, Siegler (1988) suggested that arithmetic problems may trigger additional solution strategies, if an associated result is not immediately available. Although this option was postulated, in particular, to account for errors and solution times in children (Lemaire & Siegler, 1995), it is not unlikely that it is also relevant for university students (see LeFevre, Bisanz, et al., 1996; LeFevre, Sadesky, et al., 1996). Therefore, it is possible that the negativity observed over the right hemisphere reflects such different strategies as decomposition or problem reordering. Another possibility, which seems even more likely, follows from neuropsychological case reports and brain imaging studies. These showed that representations and operations triggered by mental calculation problems are widely distributed over both hemispheres (see, e.g., Dehaene, 2000). Among other things, it was found that exact calculation is bound to neural networks in the left hemisphere whereas number comparison tasks and magnitude estimations recruit cell assemblies in the right hemisphere. Although the surface topography of an ERP effect gives only limited information about the location of its generator(s), it is nevertheless striking that the problem-size effect observed in the present study had its maximum amplitude over the right hemisphere, exactly over those brain areas that have been found as essential for magnitude estimation in production tasks (Kiefer & Dehaene, 1997). In a recent study (Jost, Beinhoff, Hennighausen, & Rösler, 2003), we compared small and large problems in a production task and also found, among other things, an additional negativity for large problems over right temporal sites. This suggests that large problems compared to small ones invoke an additional magnitude estimation routine, possibly because the associations between operands and solutions are less well established for larger problems, and the accessed solution is validated by an additional magnitude check.

In summary, our data clearly support the position that the problem-size effect is not only due to differences in the activation of the correct result but rather that it is caused by different solution strategies, too.

Evaluation of Semantic Meaning

As outlined in the Introduction, in the semantic domain, the incongruency and the constraint manipulation should affect the same functional mechanism, and, therefore, both should become manifest in a topographically equivalent and typical N400 effect. In more detail: Nonfitting words should evoke an N400 effect with about the same amplitude in any case, because they are expected to receive no preactivation at all, neither with a highly nor with a moderately constraining sentence fragment. In addition, a congruent ending after a moderately constraining sentence should also evoke at least a small N400 effect when compared with a congruent ending after a highly constraining sentence, because the former is assumed to be less activated by the priming context than the latter.

Both predictions were nicely confirmed by the data. Incongruent endings evoked a typical N400 effect with a substantial amplitude and a clear centro-parietal maximum after both highly and moderately constraining sentences. Likewise, congruent completions to moderately constraining sentence fragments in comparison to congruent completions to highly constraining fragments evoked a negativity, which had very similar features to the typical N400 effect. However, not only congruent but also incongruent endings of moderately constrained sentences evoked an N400 effect when compared with incongruent endings of highly constrained sentences. This is an unexpected result, because incongruent endings should not be primed, neither after highly nor after moderately constraining sentences. An explanation for the amplitude variation of incongruent endings in the high and moderate constraint condition might be sought in differences of the sentence material. Frequency of usage that is known to normally have an effect on N400 amplitude (e.g., Van Petten & Kutas, 1990) can be excluded because it was successfully controlled in the present study. However, several studies have shown that concreteness of a word may have an effect on the ERPs between 300 and 500 ms, too (Kounios & Holcomb, 1994; West & Holcomb, 2000). Holcomb, Kounios, Anderson, and West (1999), for example, found that incongruous concrete final words in sentences elicited a more negative potential than incongruous abstract words. This effect covers the typical N400 time window and extended until 800 ms with an anterior maximum. In the present study, constraint had a more pronounced and prolonged effect on incongruent than on congruent sentences, and this effect had a tendency to be more pronounced over anterior than over posterior sites (see Figure 1). At first glance, this result pattern bears similarities to the Holcomb et al. study. To address this issue, we assessed the concreteness of the target words in an additional, follow-up study (see Method section). Differences in concreteness ratings were found between targets in highly and moderately constraining sentences. These were rather small, especially for the targets in incongruent sentences (3.70 in the high constraint vs. 3.64 in the moderate constraint condition), but nonetheless significant. However, taking these effects as valid, the ERPs should show the reversed pattern, that is, a more negative amplitude for the high constraint condition with a slightly more concrete sample of words. Thus, there is no clear-cut evidence for relating the constraint effect to differences in word concreteness. The only post hoc explanation we can offer at present is that our procedure was somewhat different to previous studies. For example, in Kutas and Hillyard (1984), all sentence fragments were completed by a congruent word, that is, all were meaningful, whereas in our study, sentences could either end with a congruent or a completely incongruent word. This may have caused different expectancies in that in our study, semantic integration processes may have been postponed to the very end of the sentence. Given that moderately constraining sentences may preactivate concepts to a lesser degree which are related to the sentence fragment as such (without the final word), then the system had to do more semantic integration work after a moderately constraining sentence at the time when an incongruent final word was presented.

Nevertheless, it must be stressed that both manipulations, contextual constraint and cloze probability, evoked ERP effects with very similar features characteristic for a typical N400 effect, that is, a peak latency between 300 and 500 ms and a symmetrical maximum amplitude over central to parietal cortex areas. Although the statistics indicate that the topographies of the two effects are different, that is, the N400 effect evoked by the constraint manipulation extended somewhat more to anterior scalp sites, this difference is very subtle. Thus, not too much weight should be given to the significant interaction found with z scores (see Urbach & Kutas, 2000), all the more so as the differences are far less pronounced than the differences observed in the arithmetic condition between the incongruency and the problem-size manipulation. Therefore, we tend to the conclusion that the constraint and the congruency effect are both caused by functionally comparable differences in semantic relatedness and expectancy.

Arithmetic versus Semantic Incongruencies

The characteristic features of the incongruency N400 effect were highly similar in the arithmetic and the semantic condition: Onset latency was almost the same for both materials, the maximum amplitude was in general smaller and peaked earlier in the arithmetic than in the semantic condition, but most important, the topography was virtually the same. Differences in latency and amplitude stayed within the limits observed for N400 effects in the past and were also found in the Niedeggen et al. (1999) study.

Thus, it can be concluded that the same functional mechanism was triggered by the incongruency manipulation in the one and the other domain. "Additional semantic integration" is the broadest theoretical construct that has been suggested to explain N400 effects as evoked by less expected words in sentences, by unprimed words in lexical decision tasks, or by incorrect solutions of multiplication problems. In each case, unexpected semantic elements, which do not fit with the preceding context of meaning, can be assumed to need extra effort to stabilize the system with a coherent semantic meaning. And because the amplitude of the incongruency N400 seems to be inversely proportional to the level of preactivation that an item had received from the previous context (Khader, Scherag, Streb, & Rösler, in press; Niedeggen et al., 1999; Rösler, Streb, & Haan, 2001), we think that the N400 amplitude is functionally related to activation spread in memory networks.

As in previous studies, the arithmetic N400 effect was also followed in the present study by a small but significant positivity. Niedeggen and Rösler (1999; Niedeggen et al., 1999) had found such a biphasic ERP for the arithmetic incongruency manipulation, too. In these studies, the amplitude of the positivity increased monotonically with the size of the numerical distance between incorrect and correct solutions, and it was always larger for table-unrelated than table-related incorrect solutions. The positive deflection had features typical of a P300 component, which is assumed to indicate the amount of "surprise" or "context updating" (Donchin & Coles, 1988; Johnson, 1986). Therefore, the positivity following the arithmetic N400 effect was interpreted by Niedeggen and Rösler as a correlate of the implausibility of an incorrect result.

In the present study, such a positivity was observed not only after incorrect arithmetic equations but also after incongruent words completing highly constraining sentence fragments. Following the interpretation that the positivity reflects the amount of "implausibility," one has to conclude that an incongruent word after moderately constraining sentence fragments was experienced as less implausible than an incongruent word completing a highly constraining sentence fragment, whereas incongruent arithmetic solutions are experienced as equally implausible after both small and large problems. This conclusion seems reasonable, because linguistic constraints rarely predict a single semantic alternative that is unambiguously true in the strict sense of logic as is the case with a solution of an arithmetic equation. Linguistic alternatives are more or less plausible and even a very odd completion of a moderately constraining sentence fragment may make some sense, as do modern lyrics. So, a correlate of "surprise" should indeed be less prominent for nonfitting words after sentences with moderate contextual constraints.

Conclusion

The present study suggests two causes for the problem-size effect as revealed by response time measurements in arithmetic verification tasks: First, because of the latency difference of the incongruency N400 effect, it must be assumed that activation accumulates faster for small than for large problems. This agrees with the interference model of Campbell (1995). Second, the additional negativity over the right hemisphere, which became apparent in the difference ERPs, suggests that large problems invoke rechecking or magnitude estimation routines, which are not triggered by small problems. This supports explanations of the problem-size effect as suggested by Dehaene and Cohen (1995), but it is also compatible with the assumption that calculation procedures are invoked more often if fact retrieval suffers from too much interference (Siegler, 1988).

REFERENCES

Anderson, J. E., & Holcomb, P. J. (1995). Auditory and visual semantic priming using different stimulus onset asynchronies: An event-related brain potential study. *Psychophysiology*, *32*, 177–190.

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

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

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

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

Barrett, S. E., & Rugg, M. D. (1989). Event-related potentials and the semantic matching of faces. *Neuropsychologia*, *27*, 913–922.

Bentin, S., McCarthy, G., & Wood, C. C. (1985). Event-related potentials, lexical decision and semantic priming. *Electroencephalography and Clinical Neurophysiology*, *60*, 343–355.

Blankenberger, S. (2001). The arithmetic tie effect is mainly encoding-based. *Cognition*, *82*, B15–B24.

Brown, C., & Hagoort, P. (1993). The processing nature of the N400: Evidence from masked priming. *Journal of Cognitive Neuroscience*, *5*, 34–44.

Brysbaert, M. (1995). Arabic number reading: On the nature of the numerical scale and the origin of phonological recoding. *Journal of Experimental Psychology: General*, *124*, 434–452.

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

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

Campbell, J. I. D. (1995). Mechanisms of simple addition and multiplication: A modified network-interference theory and simulation. *Mathematical Cognition*, *1*, 121–164.

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

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

Campbell, J. I. D., & Oliphant, M. (1992). Representation and retrieval of arithmetic facts: A network-interference model and simulation. In J. I. D. Campbell (Ed.), *The nature and origins of mathematical skills* (pp. 331–364). Amsterdam: Elsevier.

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

Centre for Lexical Information (1995). The CELEX lexical database (Release 2 ed.). Nijmegen: Centre for Lexical Information. Max Planck Institute for Psycholinguistics.

Chwilla, D. J. (1996). *Electrophysiology of word processing: The lexical processing nature of the N400 priming effect*. Zutphen, The Netherlands: Koninklije Wöhrmann.

Collins, A. M., & Loftus, E. F. (1975). A spreading activation theory of semantic processing. *Psychological Review*, *82*, 407–428.

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

Dehaene, S. (2000). Cerebral bases of number processing and calculation. In M. Gazzaniga (Ed.), *The new cognitive neurosciences* (pp. 987–998). Cambridge, MA: MIT Press.

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

Dehaene, S., Dupoux, E., & Mehler, J. (1990). Is numerical comparison digital? Analogical and symbolic effects in two-digit number comparison. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 626–641.

Donchin, E., & Coles, M. G. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences*, *11*, 357–427.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*, 469–495.

Fischler, I., & Bloom, P. A. (1979). Automatic and attentional processes in the effects of sentence contexts on word recognition. *Journal of Verbal Learning and Verbal Behavior*, *18*, 1–20.

Greenhouse, W. W., & Geisser, S. (1959). On methods in the analysis of profile data. *Psychometrika*, *24*, 95–112.

Hennighausen, E., Heil, M., & Rösler, F. (1993). A correction method for DC drift artifacts. *Electroencephalography and Clinical Neurophysiology*, *86*, 199–204.

Holcomb, P. H., Kounios, J., Anderson, J. E., & West, C. (1999). Dual coding, context availability, and concreteness effects in sentence comprehension: An electrophysiological investigation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 721–742.

Holcomb, P. J., & McPherson, W. B. (1994). Event-related potentials reflect semantic priming in an object decision task. *Brain and Cognition*, *24*, 259–276.

Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371–375.

Johnson, R. (1986). A triarchic model of P300 amplitude. *Psychophysiology*, *23*, 367–384.

Jost, K., Beinhoff, U., Hennighausen, E., & Rösler, F. (2003). Facts, rules, and strategies in single-digit multiplication: Evidence from event-related brain potentials. Manuscript submitted for publication.

Khader, P., Scherag, A., Streb, J., & Rösler, F. (2003). Differences between noun and verb processing in a minimal phrase context: A semantic priming study using event-related brain potentials. *Cognitive Brain Research*, *17*, 293–313.

Kiefer, M., & Dehaene, S. (1997). The time course of parietal activation in single-digit multiplication: Evidence from event-related potentials. *Mathematical Cognition*, *3*, 1–30.

Kiefer, M., & Spitzer, M. (2000). Time course of conscious and unconscious semantic brain activations. *NeuroReport*, *11*, 2401–2407.

Kounios, J., & Holcomb, P. J. (1994). Concreteness effects in semantic processing: ERP evidence supporting dual-coding theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 804–823.

Kutas, M., & Federmeier, K. D. (2000). Electrophysiology reveals semantic memory use in language comprehension. *Trends in Cognitive Sciences*, *4*, 463–470.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*, 203–205.

Kutas, M., & Hillyard, S. A. (1983). Event-related brain potentials to grammatical errors and semantic anomalies. *Memory and Cognition*, *11*, 539–550.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature*, *307*, 161–163.

Kutas, M., & Van Petten, C. (1988). Event-related brain potential studies of language. In P. K. Ackles, J. R. Jennings, & M. G. H. Coles (Eds.), *Advances in psychophysiology* (Vol. 3, pp. 139–187). Greenwich, CT: JAI Press.

Kutas, M., & Van Petten, C. (1994). Psycholinguistic electrified: Event-related brain potential investigations. In M. A. Gernsbacher (Ed.), *Handbook of pycholinguistics* (pp. 83–143). San Diego: Academic Press.

Kutas, M., Van Petten, C., & Besson, M. (1988). Event-related potential asymmetries during the reading of sentences. *Electroencephalography and Clinical Neurophysiology*, *69*, 218–233.

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

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

Lemaire, P., & Siegler, R. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, *124*, 83–97.

McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: An ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology*, *62*, 203–208.

McCloskey, M., Harley, W., & Sokol, S. M. (1991). Models of arithmetic fact retrieval: An evaluation in light of findings from normal and brain-damaged subjects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 377–397.

McPherson, W. B., & Holcomb, P. J. (1999). An electrophysiological investigation of semantic priming with pictures of real objects. *Psychophysiology*, *36*, 53–65.

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: Comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 46–60.

Niedeggen, M., & Rösler, F. (1999). N400 effects reflect activation spread during retrieval of arithmetic facts. *Psychological Science*, *10*, 271–276.

Niedeggen, M., Rösler, F., & Jost, K. (1999). Processing of incongruous mental calculation problems: Evidence for an arithmetic N400-effect. *Psychophysiology*, *36*, 1–18.

Nigam, A., Hoffman, J. E., & Simons, R. F. (1992). N400 to semantically anomalous pictures and words. *Journal of Cognitive Neuroscience*, *4*, 15–22.

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

Rolke, B., Heil, M., Streb, J., & Hennighausen, E. (2001). Missed prime words within the attentional blink evoke an N400 semantic priming effect. *Psychophysiology*, *38*, 165–174.

Rösler, F., Friederici, A., Pütz, P., & Hahne, A. (1993). Event-related brain potentials while encountering semantic and syntactic constraint violations. *Journal of Cognitive Neuroscience*, *5*, 345–362.

Rösler, F., & Hahne, A. (1992). Hirnelektrische Korrelate des Sprachverstehens: Zur psycholinguistischen Bedeutung der N400-Komponente im EEG. *Sprache & Kognition*, *11*, 149–161.

Rösler, F., Streb, J., & Haan, H. (2001). Event-related brain potentials evoked by verbs and nouns in a primed lexical decision task. *Psychophysiology*, *38*, 694–703.

Rugg, M. D., & Coles, M. G. H. (Eds.). (1995). *Electrophysiology of mind: Event-related brain potentials and cognition*. Oxford: Oxford University Press.

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on the scope of facilitation for upcoming words. *Journal of Memory and Language*, *24*, 232–252.

Siegler, R. S. (1988). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General*, *117*, 258–275.

Siegler, R. S., & Shrager, J. (1984). A model of strategy choice. In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 229–293). Hillsdale, NJ: Erlbaum.

Sokol, S. M., McCloskey, M., Cohen, N. J., & Aliminosa, D. (1991). Cognitive representations and processes in arithmetic: Evidence from the performance of brain-damaged patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 355–376.

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

Urbach, T. P., & Kutas, M. (2002). The intractability of scaling scalp distributions to infer neuroelectric sources. *Psychophysiology*, *39*, 791–808.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory and Cognition*, *18*, 380–393.

Van Petten, C., & Rheinfelder, H. (1995). Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, *33*, 485–508.

West, C., & Holcomb, P. J. (2000). Imaginal, semantic, and surface-level processing of concrete and abstract words: An electrophysiological investigation. *Journal of Cognitive Neuroscience*, *12*, 1024–1037.

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.

(Received August 19, 2002; Accepted June 4, 2003)