Proximity and Precedence in Arithmetic

David Landy (corresponding author)

Department of Psychology

University of Richmond

Richmond, VA 23226

dlandy@richmond.edu

Robert L. Goldstone

Department of Psychological and Brain Sciences

Indiana University, Bloomington, IN 47408 USA

rgoldsto@indiana.edu

Word Count: 7,481

Short title: proximity and arithmetic

Abstract

How does the physical structure of an arithmetic expression affect the computational processes engaged in by reasoners? In handwritten arithmetic expressions containing both multiplications and additions, terms that are multiplied are often placed physically closer together than terms that are added. Three experiments evaluate the role such physical factors play in how reasoners construct solutions to simple compound arithmetic expressions (such as "2 + 3 x 4"). Two kinds of influences are found: First, reasoners incorporate the physical size of the expression into numerical responses, tending to give larger responses to more widely spaced problems. Second, reasoners use spatial information as a cue to hierarchical expression structure: more narrowly spaced sub-problems within an expression tend to be solved first, and tend to be multiplied. Although spatial relationships besides order are entirely formally irrelevant to expression semantics, reasoners systematically use these relationships to support their success with various formal properties.

Keywords: symbolic reasoning, mathematical cognition, embodied cognition

INTRODUCTION

One of the central challenges facing the cognitive study of mathematical reasoning is symbolic interpretation: how do people use symbol systems as carriers of meanings? In the domain of mathematics, as in other formal languages, explicit grammars specify how compound expressions are to be interpreted in terms of their basic constituents. Despite the simplicity and explicitness of these rules, numerous studies have noted that difficulties generating solutions from mathematical expressions often result from failures to correctly interpret symbolic notation (Koedinger & MacLaren, 1997; Koedinger & Nathan, 2004; Sfard & Linchevski, 1994).

Cognitive theories of abstract formal interpretation often assume that individuals follow formal logics by explicitly representing rules of combination in some internal symbolic medium, and then applying those rules to structured symbolic representations (Fodor, 1975; Marcus, 2001). In this view, the role of perception is principally to identify and represent for internal consumption individual symbols written in the external notation. If interpretation of structured notations is a result of the application of formally expressed rules, then expressions that require more, or more difficult, rules are predicted to be harder to solve than simpler expressions, but perceptual factors should only affect the transcription of individual symbols from the visual notation to an internal representation. Thus, the hierarchical structure implicit in a phrase such as “3+5x4” results from the action of a set of represented rules.

In addition to their formal properties, commonly used notational systems have many informal properties. The properties we are most interested in are those that relate pairs or sets of symbols: pairs of symbols may be similar or dissimilar, or one symbol may be larger or more salient than another, or physically close together or far apart. This paper focuses on the impact of physical spacing on arithmetic computation of simple expressions involving addition and subtraction. There are several reasons to expect spatial properties to impact arithmetic computations. First, prior work has shown that spatial properties interact with mathematical reasoning in related but distinct domains (Kirshner, 1989; Landy

& Goldstone, 2007A). Second, spatial layout affects other psychological properties of expressions, such as overall physical size and perceptual groupings. Finally, spatial properties are an obviously essential part of any physical notation system, but in mathematics in particular, layout plays an important role in constituting meaning. For instance, subscripts and superscripts depend on their spatial positions and sizes for appropriate interpretation. Even when spacing is not formally required for interpretation, conventions often govern typical spacing.

Syntax evaluation in formal languages is well-captured by rules expressed in abstract languages, as suggested by traditional cognitive theory. However, several authors have suggested that the actual process of syntactic parsing in human reasoners is often organized around visual principles, and implemented by largely visual and motor mechanisms (Endress, Scholl, and Mehler, 2007; Landy & Goldstone, 2007A, 2007B). These mechanisms are proposed to be subject to the same constraints and biases as the rest of the visual system, and to produce sharply limited kinds of grammars, consonant with the biases of the visual system. The question is not whether symbolic or visual processes are important in mathematics; clearly both are. Rather, the question is one of where and how formal grammatical interpretations occur in the adult interpreter.

It is very plausible and usually assumed in cognitive models that the role of vision is limited to symbol identification, and precedes substantive symbolic processing. This perspective suggests that physical layout, so long as it does not interfere with the identification of symbols, should have no effect on arithmetic computation. Although not essential to any particular theory, this view has generally served as a default in discussions of symbolic reasoning (e.g., Anderson, 2005; Koedinger & MacLaren, 1997; Stenning, 2002). That is, many models of mathematical reasoning, assume that the initial representation for symbolic transformation is a straightforward transduction of the presented notation. Such models do not a priori predict any result of spacing on performance.

Kirshner (1989; see also Kirshner & Awtry, 2004) found evidence that novel notations for basic arithmetic operations are learned more easily when they conform to certain spacing practices. In particular, Kirshner reports that learners more easily applied order of operations rules (e.g. multiplying and dividing before adding and subtracting in an expression) when high-precedence operations were more closely spaced. Kirshner suggests that this spacing convention closely approximates that found in typical mathematics notations and that our knowledge of operation ordering is bound to the features (proximity, in this case) that generally correspond to them. On this view, it is the regularity of physical features in the environment that leads to the connection between close spacing and multiplication. Kirshner and Awtry (2004) propose the image of a hybrid learner, able to learn declarative rules, but also (and largely separately) sensitive to statistical environmental regularities, such as visual similarities and proximities.

Landy (2007; see also Goldstone, Landy, and Son, 2009) suggests an alternative conceptualization of the role of space in formal computation. This perspective suggests that rules in symbolic environments are themselves often implemented by low-level visual-motor processes. Visual regularities are then involved in algebraic reasoning because visual processes form the primitive operations that together constitute syntactic reasoning, rather than because regularites are typically present in an environment that receives the attention of a general statistically sensitive (e.g., connectionist) learner. Landy (2007) formalized this idea in a computational model of arithmetic computation. In this model, practiced arithmetic computation is treated as an interplay between obligatory calculation processes, visualization, and pro-multiplication biases in visual attention and grouping. When people look at pairs of numbers, both their sums (LeFevre et al, 1988) and their products (Thibodeau, 1996; Rusconi et al, 2006) are automatically activated. The model assumes that similar automatic processes govern calculation in complex expressions: sub-problems that receive the most attention are most likely to activate their solutions. Individual problems are coded in terms of both perceptual and categorical features, including operands and operations, but also including vertical symmetry (a distinctive property of "tie" problems such as 7+7) and well-groupedness. These features drive the obligatory activation of sums and products.

On this account, explicit knowledge of the rules of precedence does not drive ordinary computation behaviors. Multiplications form better perceptual groups, and the signs denoting multiplication in arithmetic better attract attention. Both of these factors cause multiplications to be performed earlier than additions, which ensures that the order of operations is typically respected. Further supporting this perspective, Landy & Goldstone (2007B) demonstrated that, in addition to spacing, other properties that bias perceptual grouping (such as similarity, connectedness, and common region) also impact accuracy in algebra. These properties as such are not typically present in algebraic expressions, or at least do not seem to be correlated with formal computation order; nevertheless, they impact both visual grouping and overall accuracy in an algebra task.

Spacing of addition and multiplication signs in typical contexts

This paper comprises an empirical exploration into how variation in the spacing of arithmetic expressions involving addition and multiplication affects computation; it is worth noting briefly how such expressions are typically spaced in ecological contexts. This issue is somewhat complicated by the fact that although addition is typically represented with a + in formal contexts, there are at least four common conventions for multiplication. In algebra, multiplication is usually denoted by concatenation, as in $ax+b$, in which no operation sign is used at all, or the dot, as in $a\cdot x+b$. In handwritten and typeset arithmetic, the cross ($\times$) seems to be more typical than the dot (and concatenation is ambiguous); in computer languages, the asterisk * is frequently used to represent multiplication.

Algebraic notations tend to space multiplications substantially closer than additions. Omitting an operation sign naturally causes the operands to be placed quite close together; mathematical typesetting programs such as LaTeX also typically place the operands surrounding a dot closer together than those around a +. Thus, in algebra, multiplications are typically closely spaced, relative to additions.

In arithmetic, this pattern is not so clear-cut. The cross sign, which is used nearly universally in textbooks and in programs such as LaTeX, is generally spaced uniformly with the + sign, as in $5 \times 3 + 2$. Landy and Goldstone (2007A) reported that in handwritten expressions, cross signs predominated in arithmetic expressions, and were spaced significantly more closely than additions, and furthermore that this difference was greatest when the two appeared in the same expression. The mean difference reported by Landy and Goldstone was very small, however. Numbers surrounding addition signs were separated by an average of 9.65mm, multiplications by 9.27mm—a difference of just 0.38mm. The typical spacing of asterisks in computer programs is currently unknown, but is unlikely to be a major source of arithmetic experience for typical undergraduates in the psychology pool, the population examined here. Thus, the contexts likely to be the most typically experienced by the arithmetic reasoners studied in these experiments—textbooks and handwritten expressions—contain only very slight biases to space cross signs more closely than plus signs.

Since across the range of contexts, multiplications tend to be more closely spaced than additions, we will refer to this variety of spacing as *consistent spacing*; we will refer to spacing as *inconsistent* when additions are more narrowly spaced than multiplications, regardless of the operation sign.

Error types and measures

In the following three experiments, college undergraduates were asked to evaluate simple expressions with various physical spacings. The solutions and solution times were recorded and analyzed. Because the theoretical considerations predict errors of particular types, we also analyzed the particular kinds of errors participants made. We analyzed three types of errors: operation errors, operand errors, and precedence errors.

One frequently found error type on single operation problems (Ashcraft, 1992) is an *operation error*, in which the answer given is the correct answer to a problem that differed from the stimulus only in its operations. For instance, a response of 18 to the stimulus $3+2\times 3$ is considered an operation error, because the answer given would be correct, if the + operation were substituted by a ×.

If the given response was the correct solution to any problem with the same operations, in which one operand was different from that in the stimulus by no more than two, that response was classified as an *operand error*. In the previous example, $3+2\times 3$, a response of 12 would be considered an operand error because it is the correct response to $3+3\times 3$.

Errors were classified as *precedence errors* when the response was that that would be obtained by performing the correct operations on the correct numbers, but in the wrong order. A response of 20 to the stimulus $3+2\times 4$ would be coded as a precedence error, because $(3+2)\times 4=20$.

Most errors could be coded as one of these three types. Of the remaining errors, most appeared to be typing errors (e.g., writing 118 or 1118 for a problem whose correct result was 18). These uncoded errors were excluded from all analyses. Some errors (particularly in Experiment 2) were compatible with multiple error definitions; for instance, a response of 12 to 3+3×2 could be coded as either a precedence error (12=(3+3) ×2), or as an operand error (12=3+3×3). Analyses were performed both across all errors consistent with a particular type, and also using only those that could be uniquely classified. Results were very similar for each measure. The more inclusive measure is reported; the more restrictive measure yielded similar results, except where noted.

Across all three experiments, median response times for participants were fairly normally distributed, and could be analyzed using analyses of variance. Error rates, however, were quite low; the distribution of overall accuracy across participants did not follow a Gaussian distribution. Since many participants made few or zero errors of particular types, and in particular conditions, arcsin transformed error rates did not generally approximate normal distributions either. Error patterns were analyzed using non-parametric categorical tests. Counts of each error across each stimulus type were generated for each participant. The effect of condition on error was evaluated using a Wilcoxon signed-rank test. To evaluate the selective influence of consistency, McNemar's test was used. In Experiment 2, which has consistent, neutral, and inconsistent spacing conditions, a nominal value was generated for each error type and participant, based on whether error frequency was ordered across the three consistency conditions: error(consistent trials) ≤ error(neutral trials) ≤ error (inconsistent trials). This measure is appropriate for McNemar's test because it generates binary values.

Spacing and simple arithmetic

Experiment 1 explores the role that contrastive spatial information plays in the evaluation of single computations; Experiments 2 and 3 extend these phenomena to multi-term problems.

Neither the default rule-based model nor the perspective that computation rules are composed of learned perceptual biases makes any predictions in the single-term arithmetic case. The former model predicts no influence of spacing on performance at all; the latter predicts such influences only when there are multiple expressions, that can lead to differences in grouping or computation order. Nevertheless, there are sound theoretical reasons to expect differences in computation in this case, related to both the physical and the internal representation of numbers and number facts.

As mentioned earlier, although the signs used here (typewritten + and × signs) are most typically nearly uniformly spaced, across the range of notations and contexts, multiplication signs are more closely spaced than addition signs. If reasoners incorporate the typical relative spacing of an operation (as opposed to the typical spacing of an operation sign) into their representation of the sign, then narrowly spaced addition problems might tend to be confused with multiplication, and vice versa, leading to increased operation errors when problems are consistently spaced.

In contrast, metaphor theory (Lakoff & Nuñez, 2000) asserts that numerosity is often processed via metaphorically related representations of physical length. On this view, a representation of perceived physical length might affect arithmetic judgments, such that large spaces would tend to be conflated with large numeric magnitudes. Estimates of numerosity have been shown to be subject to this sort of size-congruity effect (Henik & Tzelgov, 1982, Choplin & Logan, 2005; Fitousi & Algom, 2006), and to affect spatial judgments (de Hevia et al, 2006).

Along these lines, several researchers (e.g., Hubbard et al, 2005) have suggested that numbers are represented along a log-compressed linear mental space, such that larger numbers are "farther" from zero than smaller numbers are; computation is interpreted as motion along that mental number line. McCrink, et al (2007) report systematic errors in addition and subtraction computation, consonant with this suggestion, which they attribute to "operational momentum." Additions tended to be overestimated, subtractions underestimated. This is predicted by the mental number line account, if when transforming a number by moving through a representation space, people tend to over adjust—to move too far.

In our case, one might expect reasoners to use the physical spacing of the operators as an implicit cue to the distance along the number line that they should "move" when computing a value. In experiment 1 the product of two numbers tends to be larger than the sum of those same numbers. This consideration predicts that people would be more likely to add when operations are narrowly spaced, and to multiply when they are widely spaced, predicting more rather than fewer operation errors when problems are consistently spaced. Similarly, this hypothesis, which we will call the *longer is larger* hypothesis predicts that people would generate a response that is slightly too large, or slightly too small, e.g., computing 15 for the narrowly spaced 7+9.

EXPERIMENT 1

Method

Participants

48 undergraduates at the University of Illinois received partial course credit for participation in this experiment.

Procedure

Some aspects of procedure are common to all three experiments presented here. In each experiment, participants were seated in front of a computer, and shown simple arithmetic problems one at a time, in a random order unique for each participant. Symbols were presented in the LeHei Pro font on Apple Macintosh computers. All displayed symbols were 14mm wide. For narrowly spaced problems, the space between the operands was 40mm (including the operation sign). For widely spaced operations, inter-operand spacing was 100mm. The viewing distance was approximately 55cm.

Problems stayed on the screen until the participant began typing a response. Responses were typed; response times were collected from the first key-press. A 1500ms rest period followed, followed by the next stimulus. Participants were instructed to perform their calculations quickly, but the problems were self-paced. Participants received breaks every 10 minutes.

In Experiment 1, stimuli consisted of single addition or multiplication problems. Operands ranged from 3 to 8; participants solved each problem in this range twice. Once, the problem was presented with narrow spacing, as in 3+5; once, it was presented with wide spacing, as in 3   +   5. There were a total of 144 problems. The experiment took approximately 15 minutes to complete.

Results

Mean accuracy was $95.9 \pm .7\%$ across all trials. Mean correct-trial response time was 1666 ± 72ms. Using operation sign (plus or times), and spacing (narrow or wide) as categorical predictors and problem size (the larger of the two operands) as an ordinal predictor, a 2 × 2 × 6 ANOVA on median response time revealed a significant main effect of problem size ($F(5,235)=15.5$, $p<.001$) and operation ($F(1,46)=7.73$ $p<.01$). Operation and problem size also interacted, such that the response time for multiplication problems increased more with magnitude than did for addition problems ($F(5,235)=4.29$, p<.001). There was no detectable interaction between spacing and operation type, $F(1,46)=0.5$, nor a main effect of spacing $F(1,46)=.09$).

Error Analysis

The theoretical considerations did not predict an effect of spacing on errors overall, but on patterns of particular kinds of errors. To evaluate these patterns of errors, we coded the incorrect solutions. Of 282 errors, all but 71 could be uniquely identified as operation or operand errors. Of the remaining errors, nearly all appeared to be typographical errors, and were eliminated from analysis. Overall, the magnitude of errors was larger for widely spaced problems. Averaging the difference between response and the correct answer for each subject in each spacing condition revealed that responses tended to be larger than the correct value for widely spaced problems, but smaller for narrowly spaced problems ($W_+(40) = 212$, $p<.01$). Within the types, operation errors were numerically more frequent on consistent than inconsistent stimuli, $W_+(23) = 224.5$, $p<.05$. That is, in accordance with the longer is larger hypothesis, smaller problems were more frequently summed than were physically larger problems. Operand errors in addition also matched the pattern predicted by this hypothesis: errors that were within 1 or 2 of the correct result tended to be smaller when the expression was narrowly spaced than when it was widely spaced, $W_+(26) = 175.5$, $p<.05$; see Table 1. Operand errors in multiplication were not well predicted by spacing ($W_+(27) = 161$, $p\sim.49$).

(Table 1 around here)

Discussion

When operations appeared singly, spatial structure had a systematic effect on computation. Consistent spacing caused increased operation confusions. This is the reverse of what would be expected if algebraic spacing conventions were biasing interpretations, but is quite in line with the longer is larger hypothesis: bigger, wider operations are more readily interpreted as more powerful operations with bigger results. Similarly, additions were systematically biased by the size of the space, so that computed sums were larger when spaces were wide. This is, again, consistent with the overall notion that longer suggests larger. This effect appeared only in the addition operation, and not in multiplication. One possible reason why the effect of size might be selective is that while additions are frequently assumed to be computed by manipulations of the mental number line, multiplications are most often either retrieved, or computed through routine processes that may not as strongly involve magnitude representations (Smith-Chant & LeFevre, 2003). The processes involved in computing multiplications are relatively insensitive to the magnitude of the result (Harley, 1990; Whalen, 2000).

Experiment 2 explores the effect of differential spacing on computations of more complex (two operation) addition and multiplication problems presented in the horizontal format. Problems such as 3+4×7 contain at least two features making them more complex than single-operation problems. At the formal level, reasoners must parse the expression correctly (i.e., as 3+(4×7) rather than (3+4)×7). This requires the reasoner to correctly respect the order of operations. At the physical level, sub-problems within a compound expression may be spaced differently from each other. While Experiment 1 also contrasted widely and narrowly spaced problems, this within-expression variation means that one problem can be grouped together spatially, meaning that the other problem is then ungrouped. In Experiment 1, problems to be solved always formed good visual groups (since they appeared alone). In Experiments 2 and 3, it sometimes happens that problems that should be processed early form poor visual groups.

The visual account presented above (see also Landy, 2007) predicts two effects of this differential grouping: first, and in line with previous results in algebra and novel arithmetic notations, the effective order of operations rule employed in parsing is likely to be affected by grouping, such that closely spaced operations are likely to be applied first, resulting in increased order errors in inconsistent spacing conditions. Second, because reasoners have a pro-multiplication attentional bias, problems that tend to attract attention (such as well-grouped expressions) will tend to be treated as multiplications. Thus, the visual primitives account predicts that in contrast to the simple expressions used in Experiment 1, in compound expressions narrowly spaced problems will tend to be multiplied and widely spaced problems added.

Studies measuring performance on single-operation problems (see Ashcraft, 1992) typically measure values for the entire range of problems with operands from around 2 to 9; these small-value problems are heavily studied in school, and solutions have often been memorized. In order to evaluate operation order behavior, two-operation problems are, of course, necessary. However, there are many low-operand two-operation problems; Experiments 2 and 3 sample this range. Experiment 2 explores the effects of spacing on problems with very small operands (2 to 4), while Experiment 3 measures the impact of spacing on problems with a mixture of small and large numerical magnitudes.

EXPERIMENT 2

Method

Participants

55 undergraduates at Indiana University received partial course credit for participation in this experiment.

Procedure

The procedure was very similar to that of Experiment 1; only the problems evaluated by participants differed. After a brief warm-up of uniformly spaced single-operation problems, participants solved a set of 216 expressions. Each expression contained two operations, which could be either addition or multiplication. The four operation structures tested are summarized in Table 2. Every participant solved every combination of these operations over the operands *2*, *3*, and *4* (except those with three identical operands), once in each of three spacing conditions. These conditions differed in their physical layout: in the *narrow-first* condition, the left-hand terms were spaced more closely than those on the right, as in $2+3 \quad \times \quad 4$. In the *wide-first* condition, the left-hand terms were spaced more widely, as in $2 \quad + \quad 3\times 4$. Finally, in the *even* condition, both operations were identically and intermediately spaced. All symbols were 14mm wide, and were presented in the LeHei Pro font. For narrow problems, the space between the operands was 40mm (including the operation sign). For neutral, the inter-operand spacing was 50mm. For widely spaced operations, inter-operand spacing was 100mm. Notice that evenly spaced problems occupied total of 142mm, but in the uneven conditions, the total horizontal extent was 182mm. Participants were reminded of the order of operations rule, and shown an example of its application before beginning the task.

(Table 2 about here)

Results

Response Time

Figure 1 shows the mean time to first key press of correct responses in each analyzed problem condition. Median response times were computed for each participant and condition, and were analyzed with a 4 (operator order: plus-plus, times-times, plus-times and times-plus) × 3 (spacing: narrow-wide, neutral, and wide-narrow) ANOVA using operation structure and spacing as categorical independent factors. In this coding, spatial-operation consistency appears as an interaction. This interaction was significant ($F(6,324)=10$, $MSE=1.1$, $p<.0001$). As can be seen in Figure 1, the interaction was due to problems in which the order of precedence differed: plus-times and times-plus problems. These were solved more quickly when the spacing was consistent with the order of operations. For problems in the times-plus order, wide-first problems took longer than other types; for problems in the plus-times order wide-first problems were fastest. There was also a significant main effect of problem type ($F(3, 162) = 68$, *MSE*=21.4, $p<.0001$), such that, generally, times-times problems took substantially longer to solve than other problems.

To verify that the results in median response time did not result from different accuracy patterns across problems in the various conditions, an items analysis was performed on all plus-times and times-plus problems. An "item" was defined as a particular formal problem, regardless of spacing. Thus, spacing constituted a within-items condition, and operation type a between-items condition. The ANOVA revealed a significant interaction between spacing and operation structure ($F(2, 70)=33.6$, MSE=27.7, $p<.0001$). Thus, across the range of items, problems were solved more quickly when they were consistently spaced.

(Figure 1 about here)

Errors

All problems, including plus-plus and times-times trials, were included in the error analysis. In total, 1180 incorrect responses were recorded. 908 of these errors fit at least one of the three error types: operand errors, operation errors, or precedence errors. The remaining 272 unclassified errors appeared to be primarily typing errors (e.g., writing 118 for a problem whose correct result was 18). 359 errors could not be uniquely categorized (i.e., the same response could result from multiple errors) leaving 549 uniquely classifiable errors

Operation errors were explored by counting errors made by each participant on consistently, neutrally, and inconsistently spaced simple expressions. For times-times and plus-plus problems with uneven spacing, one problem is "consistent" while the other is "inconsistent." That is, for a problem such as 3+3 + 4, a response of 15 (3+3×4) would constitute an operation error on a consistent expression, while a response of 13 (3×3+4) would constitute an operation error on an inconsistently spaced expression. Participants made more errors on inconsistently spaced stimuli than either neutrally ($W_+(42) = 856$, p<.001) or consistently spaced ($W_+(43) = 899$, p<.001). Consistent and neutral error rates did not differ ($W_+(37) = 391$, p~.55). This pattern held even when considering just problems with identical operations. On times-times and plus-plus problems, more errors were made when spacing was inconsistent than when it was neutral ($W_+(31) = 484.5$, p<.001) or consistent ($W_+(31) = 455.5$, p<.001). The latter two error rates did not differ ($W_+(16) = 37$, p<.1154).

Precedence errors also generally increased as spacing grew more inconsistent. 18 participants made more precedence errors on consistent than inconsistent problems, while only 1 participant did the reverse (the remaining participants made identical numbers of errors in both spacing types). The three problem types were well separated: participants made more errors on inconsistent than neutral expressions ($W_+(19) = 152$, p<.05) and more errors on neutral than consistent expressions, ($W_+(26) = 289.5$, p<.01). The difference between inconsistent and consistent errors was also significant ($W_+(27) = 327$, p<.001). Although participants made more errors overall on inconsistent than consistent expressions (see Figure 2), the relationship between precedence and consistency was particularly strong. Precedence errors were ordered by consistency for 42 of the 55 participants, but non-precedence errors were ordered for only 23 participants. Precedence errors related more strongly to spatial consistency than non-precedence errors by a McNemar's test ($\chi^2(1) = 11.17$, p<.001).

(Figure 2 about here)

Participants made more errors overall on plus-times than on times-plus stimuli ($W_{+}(46) = 933$, p<.001), suggesting a tendency to evaluate expressions from left to right, consistent with reading order. Also, operation errors were more common with operations in a right-biased order (additions on the left and multiplications on the right), than those in a left-biased order ($W_{+}(46) = 789.5$, p<.001).

When all errors were included, participants were significantly biased toward the overestimation of widely spaced, and the underestimation of narrowly spaced operations ($W_{+}(45) = 806$, p<.01). Inspection of the error patterns indicated that operand errors could frequently also have resulted from order reversals; an influence of spacing on computation order could have accounted for the bias on result magnitude, Indeed, when only operations which could be uniquely coded were included, this bias was not significant ($W_{+}(30) = 214.5$).

Discussion

The alignment of space and precedence demonstrated previously in algebra (Landy & Goldstone, 2007) and arithmetic using an invented notation system (Kirshner, 1989), was replicated here using standard arithmetic notation. When operation precedence and spatial proximity conflicted, arithmetic computations were substantially more difficult than when they were congruent. Error analysis indicated that precedence was particularly sensitive to consistency, as had been previously reported.

Operands were also more likely to be summed when widely spaced, and to be multiplied when narrowly spaced, supporting the theory that reasoners encode information about operation spacing, and use it to select operations. These are striking errors because they reflect misperceptions of clearly presented expressions. This is predicted by the visual primitives account, because easily grouped, early performed computations tend to be multiplications. A similar account can be motivated from general statistical sensitivity to conventional notations, assuming that reasoners generalize spacing regularities by operation across symbol (that is, from the typical multiplication symbols of algebra). However, this account does not easily accommodate the pattern in the single-operation case examined in Experiment 1. In the visual account, errors depend on comparative grouping, a property that does not exist in single-operation expressions, rather than on the statistical presence of differential spacing. Since the error pattern found here is the opposite of that found in Experiment 1, which had similar notational considerations but different grouping properties, we conclude that the increase in operation errors found for widely spaced multiplications and narrowly spaced additions results from the grouping created by the differential space, rather than spacing per se.

Finally, error measures show a general bias favoring the times-plus format: participants are accurate on these expressions than on plus-times expressions, and are more likely to treat an operation as a multiplication if it appears on the left.

Experiment 3 serves primarily as a replication of Experiment 2 with operands that come from a larger range. The same hypotheses are tested, in largely the same format. The problem set presented to participants is different, however, permitting an evaluation of the particular materials employed in Experiment 2, and verifying that the results are not particular to that problem set. Since the effect of spacing on operation errors reversed direction between experiments 1 and 2, it seemed prudent to verify that the latter effect was robust to other small changes in format. Furthermore, the use of larger numbers (up to 9) provides a better window onto error patterns, because particular responses are less likely to be compatible with multiple errors. Furthermore, each participant solved a particular sub-problem (e.g., 4×2) fewer times in Experiment 3 than in Experiment 2. Finally, while in Experiment 2, half of all problems could be solved by performing either the left or the right computation first, in Experiment 3, all problems contained one addition and one multiplication; thus the order was entirely specified by the rules of precedence. On the other hand, this implies increased redundancy: participants could use the identity of the left-hand operation to constrain the right-hand operation. In Experiment 2, the operations were independent. Despite these procedural differences, the general account of computation ordering as rooted in processes of attention and grouping predicts similar resuls in experiments 2 and 3.

EXPERIMENT 3

Method

Participants

38 Indiana University undergraduates received partial course credit for participation in this experiment.

Procedure

The experiment design and procedure were identical to Experiment 2. Stimuli were similar to Experiment 2, but only the times-plus and plus-times operation structures were included, and evenly spaced stimuli were dropped. The operands systematically varied in magnitude. The middle operand was always 3 or 4. Each outer operand could be independently small (2 or 3) or large (6, 8, or 9), providing compound expressions with a range of sizes and difficulties. All problems satisfying these criteria were presented, once in each of the consistent and inconsistent spacing conditions. In all, each participant saw 200 expressions in a unique random order. The experiment took about 45 minutes to complete.

Results

Response Time

The larger of the two outside operands was used as a measure of problem size. A 2 (operation order: times-plus or plus-times) × 2 (spacing: narrow-wide versus wide-narrow) × 5 (problem size) ANOVA of participants' median correct-trial response times was performed. The analysis revealed a main effect of operation order ($F(1,37)=12.0$, $MSE=2.1$, $p=.001$; see Figure 3), such that problems that had to be computed from right to left (that is, plus-times problems), and of problem size ($F(4,148)=69.6$),$MSE=56$, $p<.001$), such that problems with larger operands took longer to solve. Consistency—the interaction between spacing and operation structure—also impacted response time ($F(1,37)=27.5$, $MSE=7.1$, $p<.001$). No other effects approached significance. In particular, the three-way interaction between spacing, structure, and problem size was not significant ($F(4,148)=.53$, $p\sim.72$).

To verify that the results in response time were not due entirely to a distribution of errors across problems of different sizes, an items analysis identical to the ANOVA reported in the previous paragraph was performed, grouping all stimulus items that represented a particular formal problem, regardless of order and spacing. The analysis confirmed a significant interaction between operation order and spacing ($F(1, 49)=14$, $MSE=147$, $p<.001$).

(Figure 3 about here)

Errors Analysis

Once again, errors were classified as operation errors, operand errors, and precedence errors. These errors made up 1001 (70%) of all 1421 recorded errors. Most of the remaining errors appeared to be "double errors," in which two errors were made on the same problem; most of the rest appeared to be typos. It should be noted that the ability to uniquely identify error types increases with the magnitude of the operands. For instance, 10 was a common response for the smallest problem tested, $2 + 3 \times 2$. This could result from a precedence error, because $(2 + 3) \times 2 = 5 \times 2 = 10$, but it could also result from an operand error, because $2 + 4 \times 2 = 2 + 8 = 10$.

(Table 3 about here)

As in Experiment 2, both precedence and operation errors were more common on inconsistently than consistently spaced stimuli (precedence errors: $W_{+}(19) = 164.5$, p<.01, operation errors: $W_{+}(35) = 566.5$, p<.001) and once again, operand errors were not ($W_{+}(29) = 193.5$; see Table 3). According to a McNemar's test, the relationship between precedence errors and consistency was greater than that between consistency and all non-precedence errors ($\chi^2(1) = 6.7$, p<.05).

(Table 4 about here)

Participants made significantly more operation errors on trials with additions on the left and multiplications on the right than the reverse ($W_{+}(35) = 440$, p<.05; see Table 4).

Finally, the longer is larger hypothesis was not supported in this experiment. Participants did not systematically undestate the value of narrowly spaced problems, nor overstate the value of widely spaced problems ($W_{+}(34) = 299.5$).

Discussion

Experiment 3 successfully replicated the primary findings of Experiment 2. Experiment 3 employed a different set of stimuli, larger operands, and a different collection of spacing and operation structures than Experiment 2, but in both cases alignment of proximity and operation order increased overall accuracy, decreased accurate-trial response times, and decreased specifically precedence and operation errors. Experiment 3 verifies that the differences in operation errors between Experiments 1 and 2 did not result from the particular selection of problems involved in Experiment 2.

In general, errors increased with the magnitude of the operands, particularly errors associated with retrieving values for memorized operations (operation and operand errors). Precedence errors were mediated by spacing, but were relatively insensitive to operand size in this study. This suggests that order of operation evaluation is executed largely independently of the calculation itself.

GENERAL DISCUSSION

Spacing plays a substantial and varied role in determining how undergraduate students solve simple arithmetic expressions. Across three studies, participants were sensitive to the relative spacing of sub-problems within an expression. Spatial information affected computation in two substantially distinct ways: at the level of individual computations, and at the level of expression structure. At the level of expression structure, people preferentially grouped terms when they were placed close together. In other words, participants were more likely to execute a calculation relatively early, and to multiply, when the participating symbols were relatively closely spaced. At the level of individual computations (Experiment 1), spacing systematically affected the direction of calculated results, so that wide spaces caused calculated response to be larger than the correct result.

Spacing affects participants' executed formal structure in computation: participants tended to calculate operations early if they were closely spaced. This by itself replicates in standard arithmetic results found by Landy and Goldstone (2007B) on an algebraic validity task, and Kirshner (1989) on an alternative-notation arithmetic task. The current results go beyond previous research in two ways: first, prior work demonstrated effects of grouping in systems (algebra, or invented notation systems) which reliably incorporate spacing information. In arithmetic, in contrast, spacing information is often neutral or misleading. Although people tend to space more tightly grouped operations very slightly more closely (Landy & Goldstone, 2007A), typeset sources such as elementary school textbooks and LaTeX-formatted documents typically do not differentially space the plus and cross signs. Nevertheless, the current results demonstrate that people incorporate spacing into operation ordering even when it is only occasionally present and generally unreliable. Secondly, while both Kirshner (1989) and Landy and Goldstone (2007B) showed that spacing influences precedence errors, the current work demonstrates additional specific behaviors that are impacted by spacing. Participants are guided by the spacing not only in the order in which they apply operators, but also in the identification of individual operations: spacing is used as a cue to operation type. The influence of this cue depends on the global structure of the embedding problems, or problem set. If spacing causes differential grouping in the larger expression (Experiments 2 and 3), then close spacing implies multiplication; when it does not (Experiment 1), widely spaced (and consequently longer) operations suggest multiplication. Neither of these error effects is readily predicted by traditional models of either single or compound arithmetic computation, nor has either been to our knowledge previously reported.

This is the first demonstration that we know of that perceived expression structure impacts sub-problem computation. Models of single-problem arithmetic (McCloskey, 1992, Dehaene & Cohen, 1995; Campbell, 1994), as well as models of multi-term computation (Anderson, 2005; Koedinger & MacLaren, 1997), generally assume that single-problem computation is independent of abstract problem structure. In the situation demonstrated here, the actual computation process itself seems to be altered by the structure of the expression in which it is embedded.

It might well be possible to accommodate spacing-structure alignment biases within a generic production-system account of mathematical reasoning. After all, these regularities do exist to some degree in handwritten expressions (and in printed expressions for some multiplication signs), and it might be supposed that a system learning the order of operations would be sensitive to such statistical regularities. Therefore, a learning system could potentially incorporate this information, but could not, in general, profit from it, because the cues guiding structure are already unambiguous in the symbols themselves. Thus, there is no good reason why systems with these regularities would be in any way superior to those that lack them. Furthermore, such a generic system provides no reason to predict these results beforehand.

An alternative account that comports more naturally with the results presented here, and provides a motivation for systems that align visual and formal properties, is that the production system learning the parsing of notations is not generic, but is itself partially implemented by a highly biased visual system. That is, we speculate that real rule learning is often (and here) accomplished largely by modally specific systems with idiosyncratic learning biases (Goldstone & Barsalou, 1998; Pothos et al, 2006; Endress et al, 2007). In this case, the bias in arithmetic derives from a general visual bias to group together proximal elements into compound 'objects.' This, in turn, suggests that visual processes come to govern, in typical cases of computation, ordering operations.

The observed tendency to add widely spaced and multiply narrowly spaced problems, only when those problems appear in compound expressions, suggesting that the calculated order of operations biases peoples' perceptions of the operators themselves. In particular, when physical spacing biases people to perform an addition operation before a multiplication operation, they end up being more likely to perceive the addition operation as a multiplication operation. This bias is somewhat reminiscent of the perceptual fluency heuristic (Jacoby & Dallas, 1981; Whittlesea & Leboe, 2003). This heuristic is grounded in the robust effect that people have an easier time perceptually identifying objects that have been presented to them earlier or are somehow more familiar. Perceptual fluency reverses the causal direction of this effect, and is thus a bias to judge that items as more familiar when they are easier to perceptually identify. When items are presented in a physical manner that makes them harder to see, by rendering them in a blurry or noisy fashion, people judge that they have not been previously exposed to the item. The current result is analogous. In both cases, when perceptual processing of an item is manipulated, people are sensitive to the resulting psychological consequences on their performance, and end up incorrectly attributing the basis for their performance consequences. These phenomena are predicted when people have simultaneous failures and successes in their metacognition. On the positive side, the reasoner is using their observed order of executed operations to infer what the operators in fact were, and they are apparently doing this in an automatic fashion. On the negative side, they are unaware that their own execution order has been influenced by a formally irrelevant factor – physical spacing. It is reasonable that people do not make the correct inference to physical spacing because (like variations in stimulus blurriness for Jacoby & Dallas, 1981), these moment-to-moment variations are not a typical feature of their environment.

Finally, Experiment 1 showed an impact on individual computation that is not obviously related to structure: errors on widely spaced problems were more likely than those on narrowly spaced problems to be larger than the correct response, both by biasing the numerical computation, and causing participants to systematically misperceive operators. This result serves as experimental confirmation of the role of physical size in literal computation. The numeric size of a calculation is apparently conflated with or inherently connected to physical size such that when physical size is large, numerical result is overestimated as well. This account is well-predicted by modern accounts of mental number representation (e.g., Hubbard et al., 2005), and by metaphorical accounts of mathematical reasoning (Lakoff & Nuñez, 2000).

Because spatial consistency affects those aspects of expressions most directly involved in symbolic literacy, the interaction between space and formal reasoning has potential methodological implications for practices in the psychology of mathematical reasoning and learning. Koedinger and Nathan (2004), for example, find that, contrary to the expectations of most educators and researchers, some story and word problems are easier for high-school students to solve than formally equivalent symbolically expressed computations. Although it does not affect their main conclusion that learning to read symbolic notation is a difficult and lengthy process, it is nonetheless telling that their symbolic expressions—which require participants to understand and apply order of operations rules—all seem to be uniformly spaced, making symbolic interpretation more difficult than it would be in at least some common notations. In general, studies of this sort do not report spacing conventions; the physical spacing must be inferred from the sample figures, which in this case use a uniformly spaced font. The current research highlights the importance in educational studies in mathematics of reporting the exact physical properties of experimental stimuli. Furthermore, experiments using symbolic stimuli that do not match the spacing productions that students themselves employ when they produce mathematical expressions may not reflect real student understanding.

Attending to the role of physical layout in formal reasoning could potentially lead to the development of formalisms that offer pedagogical improvements over neutral formats. Reasoners use space when interpreting arithmetic and algebraic expressions. One might take either of two pedagogical lessons from such a reliance: first, one might take the use of spacing as a weakness of extant formal symbol-systems (and how they are taught), and attempt to design systems that do not provide such false lures; this is the approach taken by Kirshner & Awtry, 2004. We are inclined toward the alternative approach of viewing spatial structure as a virtue, not a fault, of notational systems. If environmental spatial properties are systematically constrained to bias reasoners toward correct answers, then spatial properties facilitate interpreting and evaluating expressions, benefiting the reasoning processes that often depend on these foundational skills by freeing resources potentially involved in both. However, whether this potential advantage is or can be realized in actual mathematical reasoning using standard notation is currently unclear. Substantial research into how reasoners do or can incorporate mathematical spacing practices into reasoning will be needed to reach clear or definite implications for educational practice.

Fundamentally these results challenge the conception of human reasoning as a fundamentally abstract formal process, with errors driven by misunderstandings of formal rules and properties. Instead, visual processes with idiosyncratic biases systematically impact even such an in-principle abstract task as arithmetic. The engagement of visual features and processes indicates that formal reasoning shares mechanisms with the diagrammatic and pictorial reasoning processes with which it is often contrasted. The very word “formal” contains an implicit pun: ‘form’ may either refer to the abstract structure of a thing, or to its outward shape or appearance. We think that this pun is also implicit in the way reasoners achieve the use of formal systems like arithmetic; we frequently use the outward forms of notations—and the ways that we engage them—as proxies for inherent computational essence.

Acknowledgments

This research was funded by Department of Education, Institute of Education Sciences grant R305H050116, and National Science Foundation REESE grant DRL-0910218. We thank Lydia Nichols and three anonymous reviewers for comments and criticism.

References

Anderson, J.R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science 29*, 313-341.

Anderson, J.R. (2007). *How can the human mind occur in the physical universe?* New York: Oxford University Press.

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

Cajori, F.(1929). *A history of mathematical notations*. Lasalle, Illinois: The Open Court Publishing Co.

Campbell, J. I. D. (1994). Architectures for numerical cognition. *Cognition*, *53*, 1-44.

Choplin, J. & Logan, G. (2005). A memory-based account of automatic numerosity processing. *Memory & Cognition, 33*(1), 17.

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

de Hevia, M. D., Girelli, L, Giuseppe, V. (2006). Numbers and space: a cognitive illusion? *Experimental Brain Research 168*, 254-264.

Endress, A. D., Scholl, B. J., & Mehler, J. (2005). The role of salience in the extraction of algebraic rules. *Journal of Experimental Psychology: General*, *134*, 406-419.

Fitousi, D. & Algom, D. (2006). Size congruity effects with two-digit numbers: Expanding the number line? *Memory and Cognition, 34*(2), 445.

Fodor, J. A. (1975). *The language of thought*. NY: Crowell.

Forster, K. L. & Forster, J. C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavioral Research Methods Instruments and Computers*, *35*, 116 –124.

Goldstone, R. & Barsalou, L. W. (1998). Reuniting perception and conception. *Cognition*, *65*, 231-262.

Harley, W. S. (1990). *Associative memory in mental arithmetic*. Unpublished doctoral dissertation, Johns Hopkins University

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory & Cognition*, 10, 389-395.

Hubbard, E., Piazza, M., Pinel, P., & Dehaene, S. (2005). Interactions between number and space in parietal cortex. *Nature reviews Neuroscience, 6*(6), 435-48.

Jacoby, L.L. & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General, 3*, 306-340.

Koedinger, K. R. & MacLaren, B. A. (1997). Implicit strategies and errors in an improved model of early algebra problem solving. *Proceedings of the Nineteenth Annual Conference of the Cognitive Science Society*, 382-387. Hillsdale, NJ: Lawrence Erlbaum Associates.

Koedinger, K. R. & Nathan, M. J. (2004). The real story behind story problems: Effects of representations on quantitative reasoning. *Journal of the Learning Sciences*, *13*(2), 129-164.

Kirshner, D. (1989). The visual syntax of algebra. *Journal for Research in Mathematics Education*, *20*(3), 274-287.

Kirshner, D. & Awtry, T. (2004). Visual salience of algebraic transformations. *Journal for Research in Mathematics Education, 35*(4), 224-257.

Lakoff, G. and Nuñez, R. E. 2000: *Where mathematics come from: How the embodied mind brings mathematics into being*. New York: Basic Books.

Landy, D. *Formal Notations as Diagrams of Abstract Structure*. Unpublished doctoral dissertation, Indiana University.

Landy, D. & Goldstone, R. L. (2007A). Formal notations are diagrams: Evidence from a production task. *Memory & Cognition 35*(8).

Landy, D. & Goldstone, R. L. (2007B). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition 33*(4), 720-733.

LeFevre, J.-A., Bisanz, J., & Mrkonjic, L. (1988). Cognitive arithmetic: Evidence for obligatory activation of arithmetic facts. *Memory and Cognition, 16*, 45–53.

Marcus, G. F. (2001). *The algebraic mind: Integrating connectionism and cognitive science*. Cambridge, MA: MIT Press.

McCloskey, M., & Lindemann, M. (1992). MATHNET: preliminary results from a distributed model of arithmetic fact retrieval. In J. Campbell (Ed.), *The nature and origin of mathematical skills* (p. 365-409). Amsterdam; Elsevier.

McCrink, K., Dehaene, S., & Dehaene-Lambertz, G. (2007). Moving along the number line: Operational momentum in nonsymbolic arithmetic. *Perception and Psychophysics , 69*(8), 1324-1333.

Pothos, E. M., Chater, N., & Ziori, E. (2006). Does stimulus appearance affect learning? *American Journal* of Psychology, *119*(2) 277-301.

Sfard, A. & Linchevski, L. (1994). The gains and the pitfalls of reification: The case of algebra. *Educational Studies in Mathematics, 26*, 191-228.

Smith-Chant, B. L. & LeFevre, J. (2003). Doing as they are told and telling it like it is: Self-reports in mental arithmetic. *Memory and Cognition 31*(4), 516-528.

Stenning, K. (2002). *Seeing Reason: Image and Language in Learning to Think*. Oxford: Oxford University Press.

Thibodeau, M. H., LeFevre, J.-A., & Bisanz, J. (1996). The extension of the interference effect to multiplication. *Canadian Journal of Experimental Psychology, 50,* 393–396.

Whalen, J. W. (2000). *The Influence of Semantic Number Representations on Arithmetic Fact Retrieval*. Unpublished doctoral dissertation, Johns Hopkins University.

Whittlesea, B. W. A. & Leboe, J. P. (2003). Two fluency heuristics (and how to tell them apart). *Journal of Memory and Language, 49*(1), 62-79.

Table 1: Rate of errors by type and spacing in Experiment 1.

Error	Operation Spacing	
Type	Narrow	Wide
Operand Overestimates +	.003±.002	.009±.002
Operand Underestimates +	.002±.001	.003±.001
Operand Overestimates x	.008±.002	.011±.003
Operand Underestimates x	.015±.004	.021±.005
$+ \Rightarrow \times$	.008±.002	.012±.003
$\times \Rightarrow +$	.016±.009	.009±.005

Table 2: the operation structure presented in Experiment 2.

Operation type	Example
Plus-plus	$3+2+4$
Plus-times	$3+2\times 4$
Times-plus	$3\times 2+4$
Times-Times	$3\times 2\times 4$

Table 3: Rate of operand errors in each position in Experiment 3

Error Type	Operation Spacing		
	Narrow	Medium	Wide
Overestimate	.031±.005	.028±.004	.043±.006
Underestimate	.028±.004	.02±.003	0.021±.003

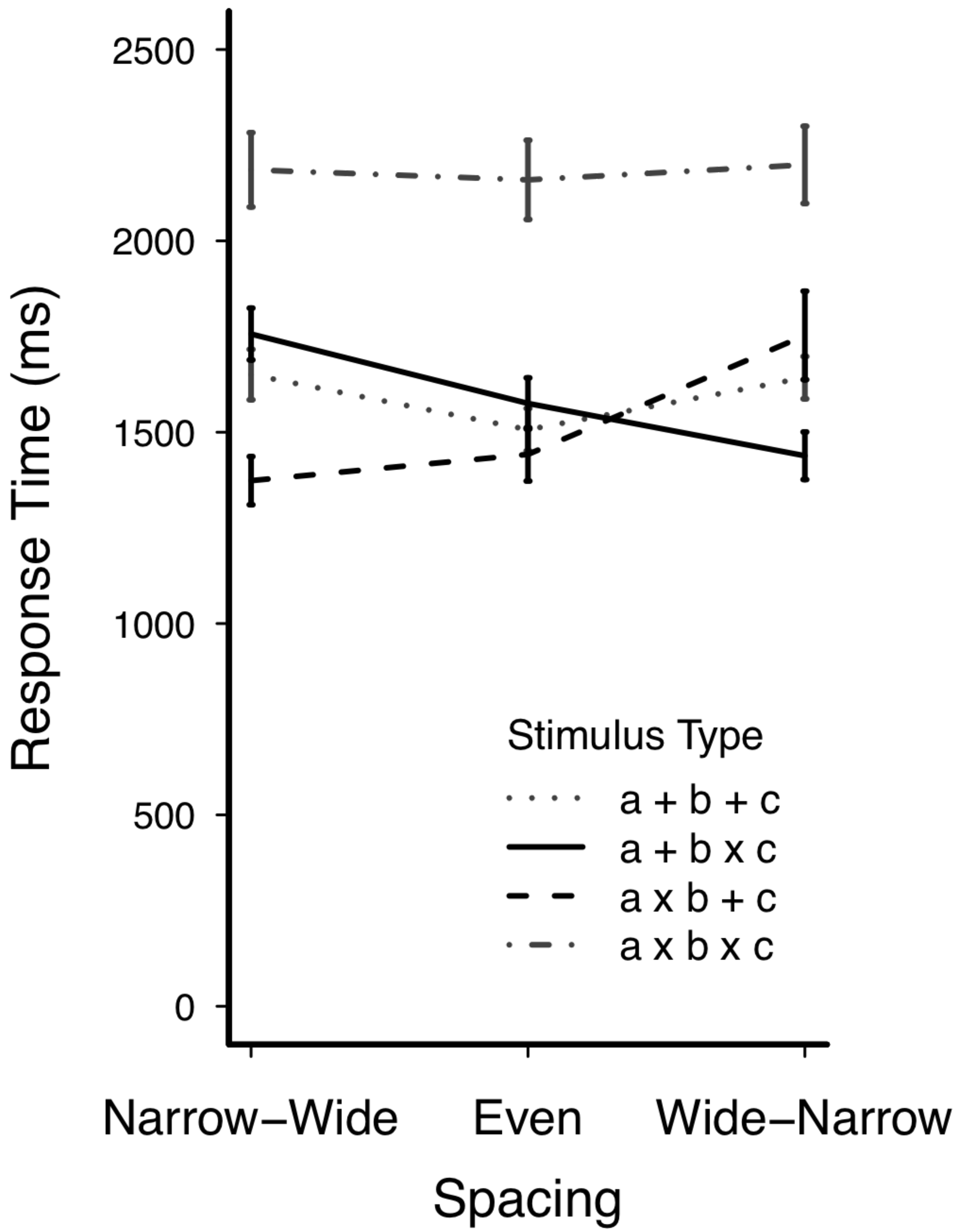
Table 4: Relationship between errors of various types and spatial consistency in Experiment 3

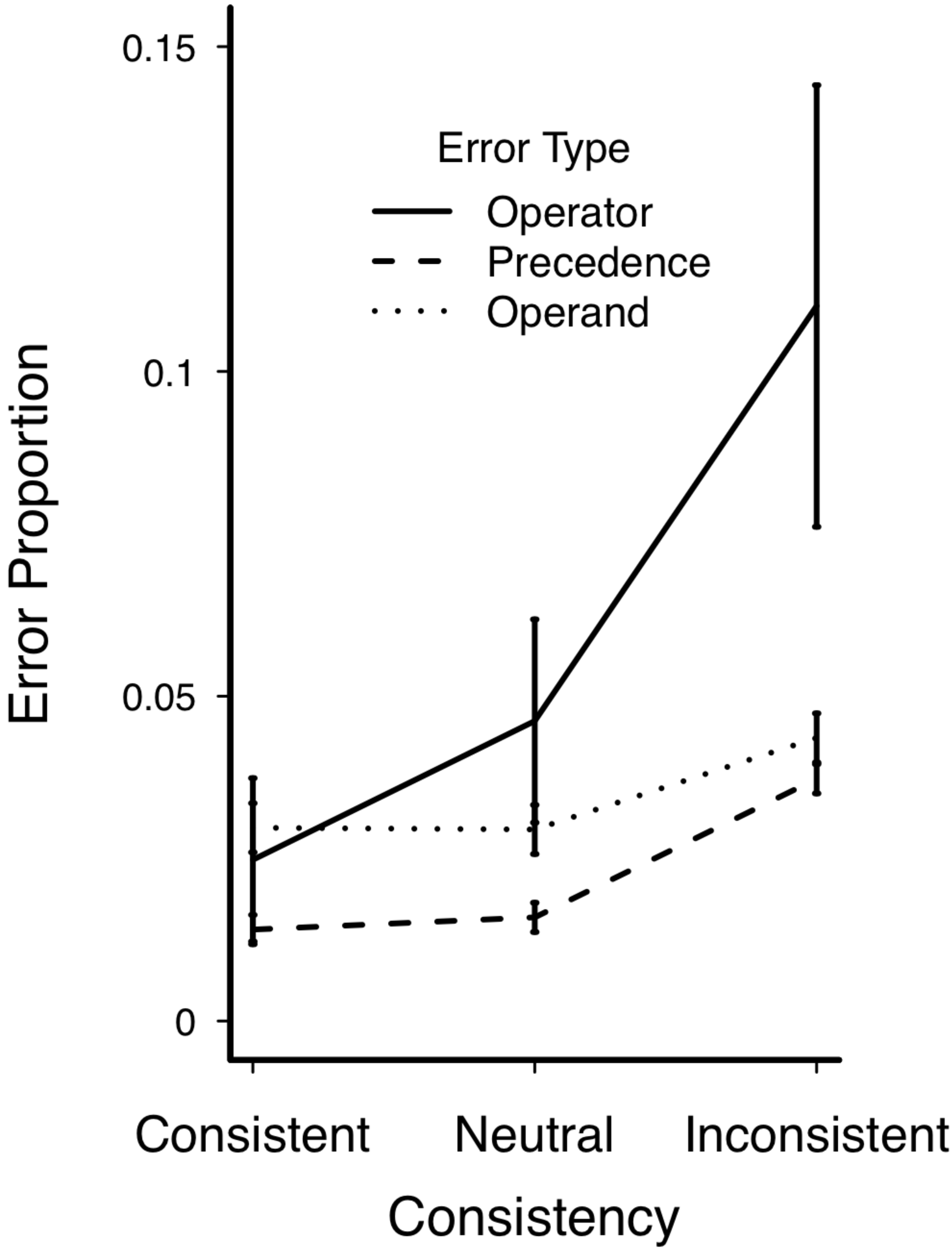
Error Measure	Stimulus Type Consistent	Inconsistent
Overall	.139±.024	.239±.047
Precedence	.030±.016	.108±.037
Operator	.014±.004	.042±.006
Operand	.034±.003	.039±.003

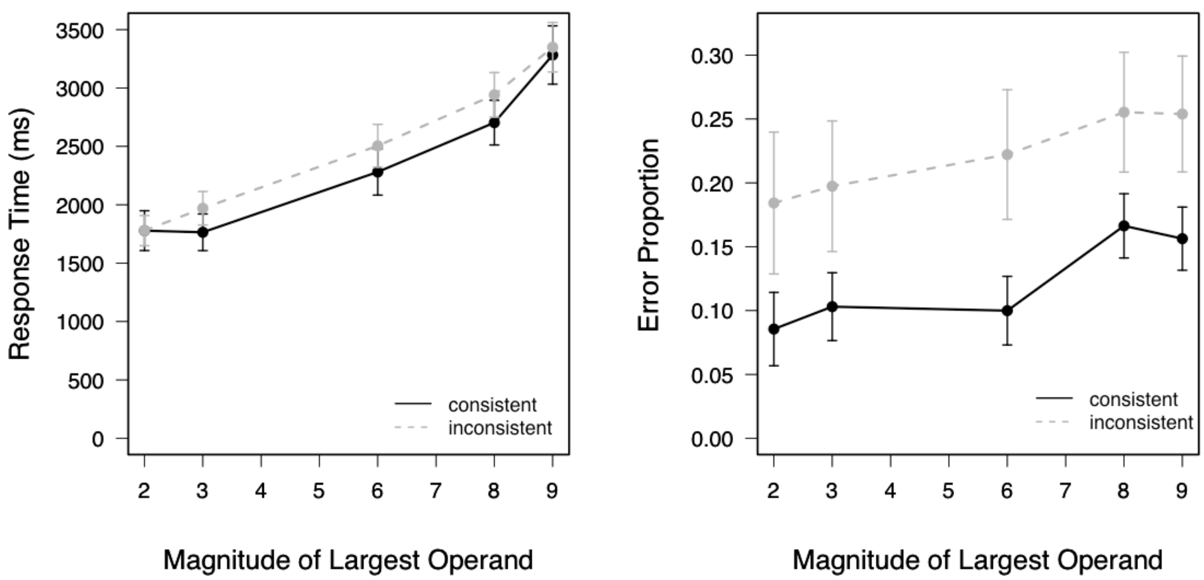
Figure 1: Mean median correct-trial response time (left) and error rate (right), for the four problem types in each spacing condition. Narrow-Wide means that the left operator is surrounded by narrow spacing, while the right operator is surrounded by wide spacing. Error bars in all cases represent one within-condition standard error.

Figure 2: Error proportion for Operand, Order, and Operator errors across several consistency levels. Error bars reflect standard errors.

Figure 3: Mean response time and error rate across operand size and consistency.


2500
2000
1500
1000
500
0
Response Time (ms)
Stimulus Type
a + b + c
a + b x c
a x b + c
a x b x c
Narrow−Wide
Even
Wide−Narrow
Spacing


0.15
0.1
0.05
0
Error Proportion
Error Type
Operator
Precedence
Operand
Consistent
Neutral
Inconsistent
Consistency


Response Time (ms)
3500
3000
2500
2000
1500
1000
500
0
2 3 4 5 6 7 8 9
Magnitude of Largest Operand
consistent
inconsistent
Error Proportion
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2 3 4 5 6 7 8 9
Magnitude of Largest Operand
consistent
inconsistent