The Alignment of Ordering and Space in Arithmetic Computation

David Landy (dlandy@indiana.edu)
Department of Computer Science
Indiana University, Bloomington, IN 47408 USA

Robert L. Goldstone (rgoldsto@indiana.edu)
Department of Psychological and Brain Sciences
Indiana University, Bloomington, IN 47408 USA

Abstract

Two experiments explore the role that spatial format has on arithmetic computations. In printed arithmetic expressions containing both multiplications and additions, terms that are multiplied are often placed closer together than terms that are added. These experiments test whether that external tendency plays a role in how reasoners construct interpretations of simple compound arithmetic expressions (such as "2 + 3*4"). Evidence is found to support two such influences: First, reasoners use spacing information when choosing an operation to apply (e.g., whether to add or to multiply two numbers). Second, reasoners also use spacing patterns to select an order in which to apply operations in compound expressions. Terms placed closer together tend to be computed sooner. Although spatial relationships besides order are entirely formally irrelevant to expression semantics, reasoners use these regular spatial relationships to support their success with various formal properties.

Keywords: symbolic processing, mathematics, embodied cognition, relational reasoning

Introduction

One of the central challenges facing the cognitive study of mathematical reasoning is symbolic interpretation: how do people use symbol strings as carriers of meanings? In the domain of mathematics, formal rules specify how terms are supposed to be interpreted. 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 read and understand symbolic notation (Koedinger & MacLaren, 1997; Koedinger & Nathan, 2004; Sfard & Linchevski, 1994). Understanding and predicting such difficulties requires that cognitive scientists learn more about how reasoners process formal notational systems.

Cognitive conceptions of abstract formal interpretation generally follow formal logics by assuming that reasoners explicitly represent rules of combination, and apply those rules to symbolic expressions (Fodor, 1975; Marcus, 2001). In this view, the role of perception is to identify and represent for internal consumption individual symbols written in the external notation. Problems that involve more, or more difficult, rules are predicted to be harder to solve, but perceptual factors should only affect the transcription of the external notation into internal symbols.

In addition to their formal properties, commonly used symbol systems have many informal properties as well: pairs of symbols may be similar or dissimilar, or one symbol may be larger or more salient than another, or bear other non-formal but readily accessible features and relations. Although these properties are not formally meaningful, it is quite likely that reasoners are sensitive to such regularities, and use them to build successful interpretations.

The major goal of this paper is to explore the role that one such property—physical spacing—plays in simple arithmetic computations. Arithmetic equations are often non-uniformly laid out on a page. For instance, in mathematical typesets and typesetting programs such as LaTeX, multiplications are automatically spaced more closely than additions. Handwritten expressions are more variable in their appearance, but our data suggests that this spacing regularity also exists in handwritten expressions (Landy & Goldstone, In Press A).

In this paper we explore three hypotheses about how reasoners might use regular addition and multiplication spacing in building arithmetic interpretations. According to the *operator feature hypothesis*, participants might be sensitive to the relational spacing regularities of various operators when they identify operator type. If an operator is widely spaced in a particular expression, therefore, it is assumed to be more likely an addition than a multiplication. This assumption may lead to people interpreting 2+4 * 7 to equal 56, if they interpret the "+" as a "*" because the close spacing leads them to perform the calculation involving the "2" and "4" before the calculation involving "4" and "7."

The *proximity-precedence alignment hypothesis* proposes that people will be sensitive to the visual hierarchy of perceptual groups present in an equation, and that this sensitivity will cause closer objects to be combined first (Kirshner, 1989; Kirshner & Awtry, 2004). Thus, the expression 2+4 * 7 is likely to be interpreted as 42, rather than (the formally correct) 30, because the spatial proximity of the threes will incline perceivers to group them formally. In this paper, we will refer to expressions as *consistent* if higher-precedence operations are more closely spaced, *inconsistent* if higher-order operations are more widely spaced, and *neutral* if evenly spaced.

Finally, because equations are typically read from left to right, we expect that expressions with products placed to the left of sums will be easier to solve (the *expression reading hypothesis*). These hypotheses will be measured against a null hypothesis of *no effect* of physical layout, which although not essential to any particular theory, has generally served as a default view in discussions of symbolic reasoning (e.g., Anderson, 2005; Stenning, 2002).

The first two hypotheses predict that arithmetic problems which violate the normal spatial relationships will be more difficult generally, but they predict different specific error patterns: the operator-feature hypothesis predicts that reasoners will make operator confusions, while the operation-order hypotheses predicts that reasoners will tend to apply the correct operations, but in the wrong order. In the following two experiments, college undergraduates were asked to compute values for simple expressions with various physical spacings. The solutions and solution times produced were used to evaluate these hypotheses.

Studies measuring performance on single-operator problems (see Ashcraft, 1992) typically measure values for the entire range of problems with operators from around 0-9; these small-value problems are heavily studied in school, and solutions have often been memorized. In order to evaluate operation order behavior, two-operator problems are, of course, necessary. However, there are many low-operand two-value problems; the goal of these experiments is to sample this range. Experiment 1 explores the effects of spacing on very low-operand problems, while Experiment 2 measures the impact of spacing on problems with a mixture of sizes.

Experiment 1

Procedure

55 Indiana University undergraduates participated in this experiment in exchange for partial course credit. Participants were seated in front of a computer, and shown simple arithmetic problems. Participants solved problems, and typed their responses into the computer keyboard. Response times were collected from the first key-press.

After a brief warm-up of single-operator problems, participants solved a set of 216 expressions. Each expression contained two operations, which could be either additions or multiplications. Every participant solved every combination of these operations over the operands *2*, *3*, and *4*, except problems with all three operands identical (errors on such problems are difficult to analyze), a total of three times, 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 * 4". In the *wide-first* condition, the left-hand terms were spaced more widely, as in "2 + 3*4". Finally, in the *even* condition, both operators were identically and intermediately spaced. The four operator structures tested are called *plus-plus, times-plus, plus-times,* and *times-times*, and are of the forms a+b+c, a*b+c, a+b*c and a*b*c, respectively. Problems were presented to each participant in a unique random order. In pre-experiment instructions, participants were asked to perform their calculations quickly, but the problems were self-paced. Each remained on the screen until the participant completed a response by pressing the return key. Participants were reminded of the order of operations rule, and given an example of its application in the instructions.

Results

Three participants failed to reach a criterion of 70% mean accuracy, and were eliminated from analysis, leaving 52 participants whose data were analyzed.

The expression reading and proximity-precedence alignment hypotheses make predictions about overall problem difficulty (measured by accuracy and correct-trial response time (RT)); the operation feature hypothesis makes predictions only about particular kinds of errors.

Half of all trials (the *plus-plus* and *times-times* trials) contained only one type of operator, and consequently have no formally defined order or consistency; these trials are not relevant to the proximity-precedence alignment hypothesis or the expression reading hypothesis, and so these trials are excluded from the analysis of overall RT and accuracy.

Response Time Spacing and operation order affected correct-trial response time. Figure 1 presents the mean RT for each problem condition. The left-hand bars reflect response times on times-plus stimuli, the right on plus-times. We analyzed overall RT on these trials with a 2-way 2x3 ANOVA using operator structure and spacing as independent categorical variables. In this coding, spatial-operator consistency appears as an interaction. As predicted by proximity-precedence hypothesis, this interaction was significant ($MSE=30.8$, $F(2,96)=31.6$, $p<0.0001$). 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. Problems in the times-plus operator format took less time to solve overall than plus-times problems ($MSE=9.96$, $F(1,49)=7.5$, $p<0.01$), as predicted by the expression reading hypothesis.

Accuracy In general, accuracy results match those found in RT, indicating that differences do not result from a speed-accuracy tradeoff. Figure 2 presents overall accuracies.

According to a 2-way 2x3 ANOVA, spacing and operation order interacted significantly ($MSE=5.1$, $F(2,102)=8.97$, $p<0.001$). Examination of the means revealed that, as predicted by the proximity-precedence alignment hypothesis, consistent problems were solved

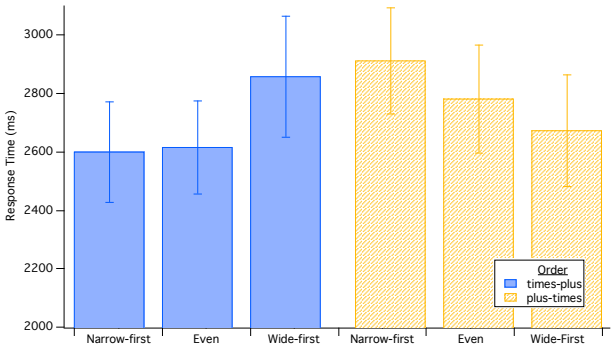

Figure 1: Mean response times (RT) in Experiment 1. Errors in this, and all following graphs are between-participant standard errors.

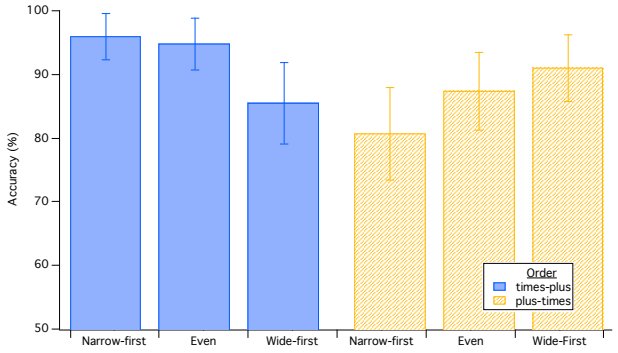

Figure 2: Mean accuracy in Experiment 1.

more accurately, and inconsistent problems less accurately than were evenly spaced problems.

Operation order also significantly affected accuracy as a main effect of the ANOVA analysis (MSE=4.5, $F(1,51)=7.8$, $p<0.01$), with times-plus problems being solved more accurately than plus-times problems, as predicted by the expression reading hypothesis.

Errors In total, 971 incorrect responses were recorded. Most of these errors uniquely matched one type of the following errors: In *operator confusion* errors, the answer given was the correct answer to a problem which differed from the stimulus only in its operators. For instance, a response of "18" to the stimulus "3+2*3" was categorized as an operator confusion error.

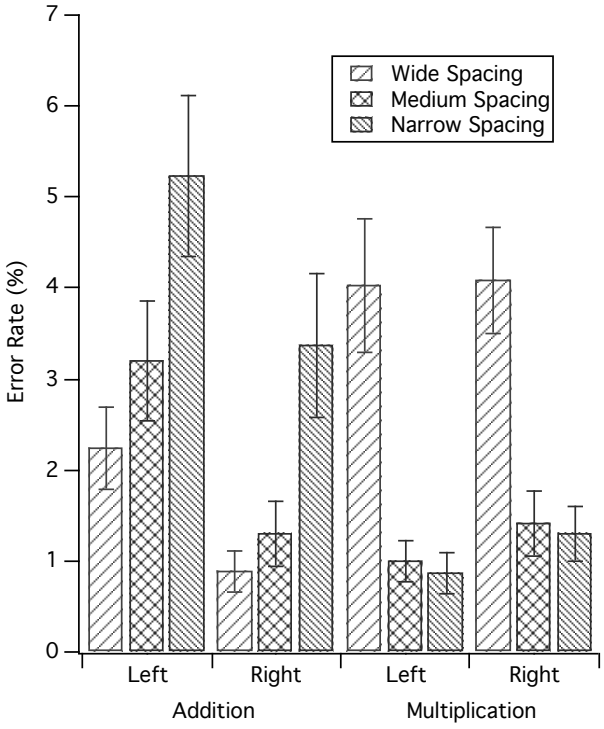

Figure 3: Rate of operator confusion errors (in percentage of all responses), for each operator.

Incorrect responses which were correct solutions to a problem with one operand different by exactly one from the actual stimulus were called *operand* errors. Thus, for the stimulus "2*3*3," the response "27" was labeled an operand error, since it is the correct solution to the problem "3*3*3."

Precedence errors occur when the result was compatible with performing the correct operations on the correct numbers, but in the wrong order. A response of "20" to the stimulus "3+2*4" would be coded as a precedence error, since $(3 + 2) * 4 = 20$.

549 of all errors could be uniquely classified into one of these three types. Of the remainder, most were typographical errors, or fit into more than one of the three primary categories. In what follows, we only analyze the uniquely classifiable errors.

The proximity-precedence alignment hypothesis predicts an impact of spatial consistency specifically on precedence errors. To evaluate this prediction, operand and precedence errors were analyzed together, using a repeated measures analysis, in which the two measures were the per-individual rate of each type of error. This measure was used in a 2-way ANOVA using error frequency as the dependent measure, with error type and item consistency as factors. There was a significant main effect of consistency, with more errors on inconsistent, and fewer on consistent trials (MSE=61.3, $F(1,51)=9.2$, $p<0.01$). The interaction term was also significant: consistency had a larger effect on precedence than on operand errors (MSE=43.7, $F(1,51)=5.1$, $p<0.05$).

The operator feature hypothesis predicts not just an increase in errors due to inconsistent spacing, but that relative spacing will specifically affect individual operator identification. To evaluate the operator feature hypothesis, the data were recoded by individual operations, and the effective operator was determined for each correct response and operator error. In the example given earlier, the effective operations on the right and left would both be coded as *times*, because the result (18) is consistent with multiplying 3*2*3. The operator feature hypothesis predicts that more widely spaced operations will be more likely to be treated as additions. The expression reading hypothesis predicts that problems on the right-hand side are more likely to be treated as additions.

Figure 3 presents the proportion of trials for which responses were operator confusions, broken down by spacing, operator position, and operation. A 3-way ANOVA using error rate as the dependent measure, and operation (addition or multiplication), operator position (left or right), and spacing (narrow, medium, and wide) as factors revealed the interaction between spacing and operator predicted by the operator feature hypothesis (MSE=1.2, $F(2, 102)=18.4$, $p<0.001$); operator confusions were more frequent with additions on the left and multiplications on the right. The expression reading hypothesis predicted an interaction between operator and spacing, with more confusions on

additions when they appeared on the left, and on multiplications when they appeared on the right. This interaction was significant (MSE=.50, $F(1,50)=9.5$, $p<0.01$).

Discussion

All three hypotheses were supported in this experiment. When operator precedence and spatial proximity conflicted, arithmetic computations were substantially more difficult. Error analysis indicated that precedence was particularly sensitive to consistency. Operands were also more likely to be summed when widely spaced, and to be multiplied when narrowly spaced, supporting the hypothesis that reasoners encode information about operator spacing, and use it to interpret either symbols or operations. Finally, all measures show a general bias toward the times-plus format: participants are faster and more accurate on these problems than on plus-times problems, and are more likely to treat an operator as a multiplication if it appears on the left.

Experiment 2 serves as a basic replication of Experiment 1. The same hypotheses are tested, in largely the same format. However, while Experiment 1 used only small-number problems, Experiment 2 evaluates a range of small- and large-number problems. Larger problems are generally more difficult than smaller ones (Ashcraft, 1992); this increased difficulty might impact the role that spacing plays in either operator or order of operations judgments.

Experiment 2

Procedure

38 Indiana University undergraduates participated in this experiment for course credit. The experiment design and procedure were identical to Experiment 1. Stimuli were similar to Experiment 1, but only the times-plus and plus-times operator structures were included, and evenly spaced stimuli were dropped. The operands used were designed to systematically vary operand size, without exceeding the number of problems that could be solved in an hour-long experiment. The middle operand was always 3 or 4. The outer operand could be small (2 or 3) or large (6, 8, or 9). In all, each participant solved 200 problems in a unique random order. The experiment took 45 minutes to complete.

Results

Eight participants failed to reach a criterion of 70% mean accuracy, and were eliminated from analysis, leaving 30 participants whose data were analyzed.

The larger of the two outside operands was used as a measure of problem size, termed the *maximum operand*. The results were also analyzed separating each operand, and separating spacing and operator order, with identical results. In this experiment, stimulus consistency was coded and used as an independent measure.

Response Time A 2-way ANOVA using correct-trial response times as a dependent measure revealed a significant role of both consistency and operand size (see Figure 4B). Inconsistent trials took longer to solve than consistent ones (MSE=53.1, $F(1,29)=22.5$, $p<0.001$), and large operands yielded slower solutions (MSE=1,208, $F(1,29)=117$, $p<0.001$). The interaction was not significant (MSE=.24, $F(1,29)=.26$, $p>0.6$).

Accuracy A 2-way ANOVA over consistency and maximum operand revealed main effects of both: accuracy was higher on consistent than inconsistent problems (MSE=0.3, $F(1,29)=4.5$, $p<0.05$, see Figure 4A), and was lower on larger operand problems (MSE=4.7, $F(1,29)=38.7$, $p<0.0001$). Consistency and operand size did not interact (MS=0.01, $F(1,29)=.13$, $p>.7$). The effect of consistency on overall accuracy was slight: mean accuracy was 89.8% on consistent trials and 91.2% on inconsistent trials.

Errors Analysis Once again, errors were classified as operation errors, operand errors, precedence errors, and other errors. These errors made up 43% of all 572 recorded errors. Most of the remaining errors fit into more than one of the above categories or 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*2. This could result from an precedence error, because (2+3)*2=5*2=10, but it could also result from an operand error, because 2+4*2=2+8=10 (see Figure 4B and 4C. Unclassified errors are presented in Figure 4D). The same analyses were also performed using non-exclusive error categories (with each ambiguous error being counted once for each of its possible error types, with essentially identical results.

The impact of problem size on precedence and the order of operand errors was analyzed using a 3-way repeated measures ANOVA over maximum operand, spatial consistency, and the type of error with error frequency as the dependent measure. The results are presented in Figure 4D and 4E. This analysis revealed a main effect of problem size (MSE=.04, $F(1,29)=27.8$, $p<0.001$), and also showed that operand errors were more common overall (MSE=0.07, $F(1,29)=37.7$, $p<0.001$). Two interactions were also significant. First, operand errors were more influenced by operand size than were precedence errors (MSE=0.029, $F(1,29)=19$, $p<0.001$). Second, precedence errors were more influenced by consistency than were operand errors (MSE=0.004, $F(1,29)=4.2$, $p=0.05$). T-tests verified that precedence errors were significantly affected by consistency ($t(29)=-2.28$, $p<0.05$), but that operand errors were not ($t(29)=0.98$, $p\sim0.34$).

Operator errors were analyzed using each operator as a separate measurement; these were analyzed using a 3-way ANOVA using error rate as the dependent measure. As in Experiment 1, operators which were more closely spaced were more likely to be treated as multiplications (MSE=.15, $F(1,29)=4.4$, $p<0.05$). Operator position did not significantly affect the perceived operator (MSE=0.08,

F(1,29)=2.11, p~0.16). The effects of maximum operand and spatial consistency are displayed in Figure 4A.

Discussion

Experiment 2 successfully replicates the major findings of Experiment 1. Experiment 2 employed a different set of stimuli, larger problems, and a different collection of spacing and operator structures than Experiment 1, but in both cases spatial consistency increased overall accuracy, decreased accurate-trial response time, and decreased specifically precedence and operation errors. Analysis of Experiment 2 verifies both the operation feature hypothesis and the precedence hypothesis.

In general, errors increased with the magnitude of the operands, particularly errors associated with retrieving values for memorized operations (operation and operand errors). Errors relating to the expression structure—precedence errors, were mediated by spacing, but were relatively insensitive to operand size in this study. This is significant, since understanding a symbolic expression requires correctly determining how the individual symbols bind together, even when a value is not actually computed. This result suggests that the factor introduced here, expression spacing, may mediate that understanding more robustly than the sizes of the operands used. No evidence was found in this study favoring the expression reading hypothesis—the assumption that participants would tend to

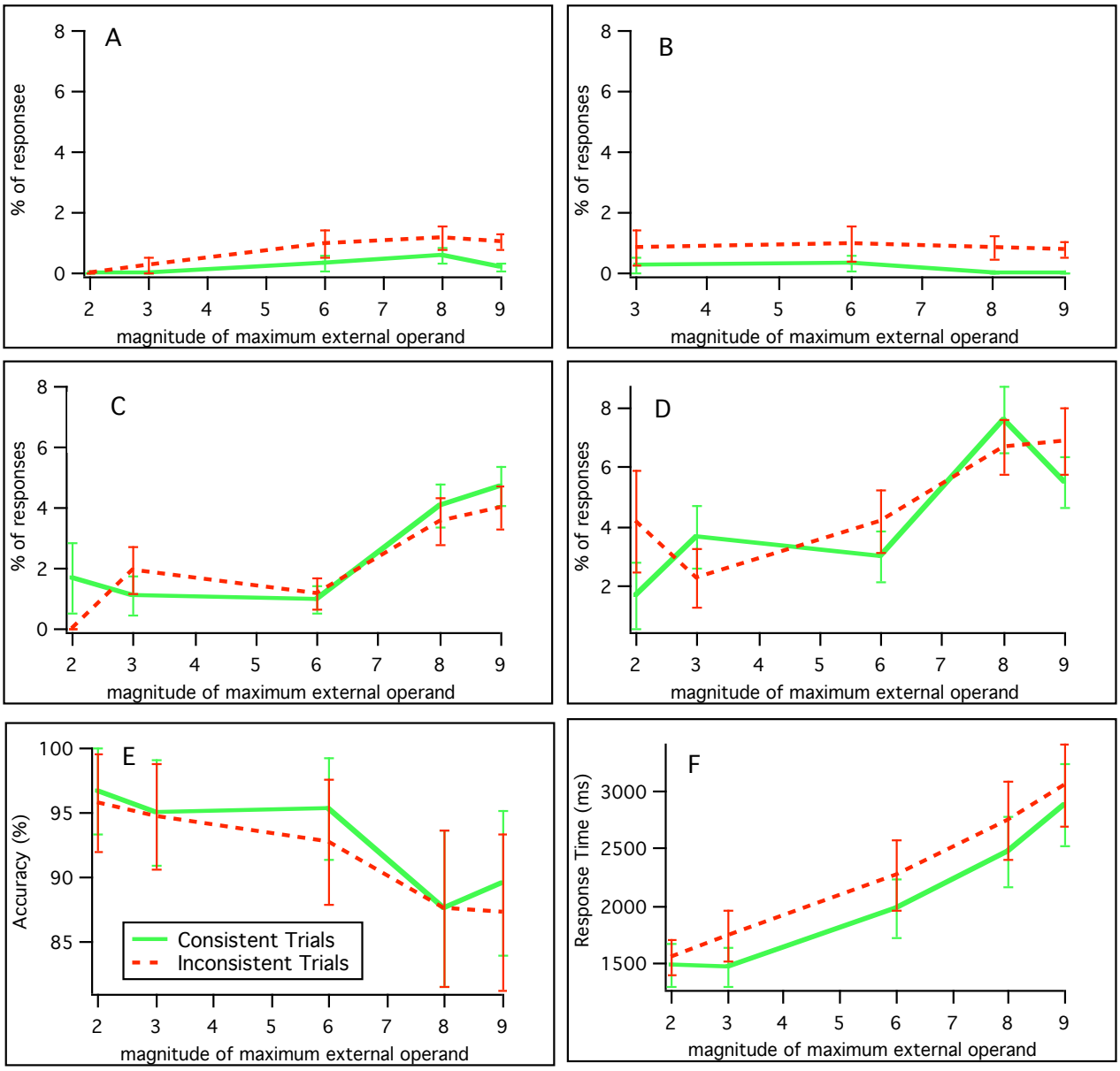

Figure 4: Effects of consistency and maximum problem operand and the frequency (in percentage of all responses of the relevant category) of operation confusion errors (A), precedence errors (B), operand errors (C), and uncategorized errors (D). Figures E and F present overall accuracy (E), and mean correct-trial response time (F).

compute left to right. Future research will determine which experiment aspects encourage operation order dependencies.

General discussion

A pair of studies explored how undergraduates use formally irrelevant spatial information to solve compound arithmetic expressions involving addition and multiplication. In both studies, participants were sensitive to the statistical tendency of multiplications to be more closely spaced than additions, and demonstrated this sensitivity in several ways. First, problems with spacing consistent with the operator structure were solved more successfully and more quickly than problems in which spacing was neutral, or in which spacing mismatched formal structure. Participants tended to multiply numbers that were more closely spaced, and add those with more space between them. Even when the correct operation was performed, spacing impacted the likelihood that computations would be performed in the correct order. Finally, in one experiment, participants were more successful when expressions were written with the higher-order operation on the left.

From an algorithmic perspective, it is interesting that precedence order was affected by spacing even when the correct operations were performed. One might have supposed that statistical regularities about spacing would impact operator recognition, but that once the terms had been correctly identified the rule system would determine the correct order (as in Anderson, 2005). It seems instead that the processes that infer structure in formal notations are also sensitive to those that represent spatial groupings. While the effects of spacing presented here are not large in absolute magnitude, the task was also extremely simple. On less well-learned tasks, in which aligning structure and interpretation is more difficult generally, semantic-spatial alignment may play a larger role (Landy & Goldstone, in press B, considers the case of elementary algebra).

Because spatial consistency affects precisely those aspects of expressions most directly involved in symbolic literacy, the interaction between space and formal reasoning has methodological implications for practices in the psychology of mathematical reasoning as well: Koedinger and Nathan (2004), for instance, 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. Their interpretation of this is that, like natural languages, mathematical formalisms take time and effort to learn, and that comprehension errors affect not just story problems but also formal arithmetic systems. 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. 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. Attending to the role of non-physical layout could make such results more informative.

Fundamentally these results challenge the conception that human reasoning with formal systems uses only the formal properties of symbolic notations, and that errors are driven by misunderstandings of those properties. Instead, people seem to use whatever regularities—formal or visual, rule-based or statistical—are available to them, even on an entirely formal task such 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 normally contrasted. In short, our research indicates that, when displayed correctly, even a sentence is worth a thousand words.

Acknowledgments

This research was funded by Department of Education, Institute of Education Sciences grant R305H050116, and National Science Foundation ROLE grant 0527920.

References

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

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

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

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.

Landy, D. & Goldstone, R. L. (in press A). Formal notations are diagrams: Evidence from a production task.

Landy, D. & Goldstone, R. L. (in press B). How abstract is symbolic thought?. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

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

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

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