Non-formal mechanisms in mathematical cognitive development: The case of arithmetic

Braithwaite, D.W.; Goldstone, R.L.; van der Maas, H.L.J.; Landy, D.H.

DOI
10.1016/j.cognition.2016.01.004

Publication date
2016

Document Version
Final published version

Published in
Cognition

Cognition 149 (2016) 40–55

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Non-formal mechanisms in mathematical cognitive development: The case of arithmetic

David W. Braithwaite [a,*], Robert L. Goldstone [b], Han L.J. van der Maas [c], David H. Landy [b]

[a] Carnegie Mellon University, Pittsburgh, PA, United States
[b] Indiana University, Bloomington, Indiana, United States
[c] University of Amsterdam, Amsterdam, Netherlands

ARTICLE INFO

Article history:
Received 6 August 2014
Revised 27 August 2015
Accepted 5 January 2016
Available online 12 January 2016

Keywords:
Mathematical cognitive development
Concrete to abstract shift
Arithmetic
Syntax
Perception
Mathematics education

ABSTRACT

The idea that cognitive development involves a shift towards abstraction has a long history in psychology. One incarnation of this idea holds that development in the domain of mathematics involves a shift from non-formal mechanisms to formal rules and axioms. Contrary to this view, the present study provides evidence that reliance on non-formal mechanisms may actually increase with age. Participants – Dutch primary school children – evaluated three-term arithmetic expressions in which violation of formally correct order of evaluation led to errors, termed foil errors. Participants solved the problems as part of their regular mathematics practice through an online study platform, and data were collected from over 50,000 children representing approximately 10% of all primary schools in the Netherlands, suggesting that the results have high external validity. Foil errors were more common for problems in which formally lower-priority sub-expressions were spaced close together, and also for problems in which such sub-expressions were relatively easy to calculate. We interpret these effects as resulting from reliance on two non-formal mechanisms, perceptual grouping and opportunistic selection, to determine order of evaluation. Critically, these effects reliably increased with participants' grade level, suggesting that these mechanisms are not phased out but actually become more important over development, even when they cause systematic violations of formal rules. This conclusion presents a challenge for the shift towards abstraction view as a description of cognitive development in arithmetic. Implications of this result for educational practice are discussed.

1. Introduction

The idea that cognitive development involves a shift towards abstraction has a long history in psychology (Gentner & Toupin, 1986; Gentner, 1988, 2003; Keil & Batterman, 1984; Keil, 1989; Piaget, 1952; Rattermann & Gentner, 1998; Vygotsky, 1962). This shift supposedly involves decreasing reliance on perceptual features and details of context, and increasing reliance on abstract features and context-free rules. In academic disciplines such as mathematics and physics, the development of expertise as a result of education and experience has also been described in terms of a shift towards abstraction (Chi, Feltovich, & Glaser, 1981; Chi & VanLehn, 2012; De Lima & Tall, 2008; Novick, 1988; Tall, 1995, 2008). However, some researchers have challenged the notion of a shift towards abstraction on both theoretical (Keil, Smith, Simons, & Levin, 1998) and empirical (Bulloch & Opfer, 2009) grounds, or even proposed that a shift in the opposite direction may occur (Simons & Keil, 1995; Varma & Schwartz, 2011).

The present study provides evidence that in the domain of symbolic arithmetic, the influence on performance of formally extraneous perceptual and contextual details increases with age and experience, suggesting that development in this domain cannot be fully characterized in terms of a shift towards abstraction. Arithmetic is an attractive domain for investigating this issue for at least two reasons. First, there exist explicit formal rules constraining correct performance in arithmetic, so it is natural to suppose that arithmetic competence consists precisely in following these rules, and that the development of such competence involves a shift towards such formal rule-governed behavior. Thus, one might expect arithmetic to be a likely domain for showing a developmental shift towards abstraction. Secondarily, arithmetic is of immense practical importance, due to its direct utility in a wide range of

* Corresponding author at: Department of Psychology, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213, United States.
E-mail address: baixiwei@gmail.com (D.W. Braithwaite).

http://dx.doi.org/10.1016/j.cognition.2016.01.004
0010-0277/

other domains as well as its foundational role in higher mathematics. Understanding the nature of learning and development in arithmetic has the potential to inform instructional design and thereby improve educational outcomes.

We focus more specifically on the evaluation of complex arithmetic expressions – that is, expressions involving multiple operations. Intuitively, such expressions could be evaluated by selecting and evaluating simple sub-expressions until a single value is reached. For example, faced with the complex expression "1+3×2," one might first evaluate "3×2" as "6," then "1+6" as "7." The processes underlying evaluation of simple arithmetic expressions (e.g. "3×2" and "1+6") are well-understood, and include counting-based strategies, calculation algorithms, and retrieval from memory (Brissiaud & Sander, 2010; Miller, Perlmutter, & Keating, 1984; Moore & Ashcraft, 2015; Shrager & Siegler, 1998; Siegler & Stern, 1998). Less well-understood are the mechanisms by which simple sub-expressions are selected and prioritized for evaluation in the first place. For instance, in the above example, how does one decide to begin by evaluating "3×2" rather than "1+3?" We first describe three mechanisms that could support such selections: syntactic parsing, perceptual grouping, and opportunistic selection. We then discuss the possible roles of these mechanisms over the course of learning and development.

1.1. Mechanisms

1.1.1. Syntactic parsing

In syntactic parsing, evaluation of complex expressions is preceded and guided by their syntactic structure, which is determined according to formal rules of syntax. For example, applying rules of operator precedence to the expression "2+7×5" would allow one to identify "7×5" (but not "2+7") as a syntactic phrase within the larger expression. This simpler sub-expression could then be evaluated directly via retrieval from memory. As another example, applying the rule for left-to-right evaluation among operators of equal precedence to the expression "25–13–3," one would identify "25–13" (but not "13–3") as a syntactic phrase, which could then be evaluated.

Consistent with such a mechanism, adults trained in arithmetic and algebra are sensitive to syntactic structure (Jansen, Marriott, & Yelland, 2003; Schneider, Maruyama, Dehaene, & Sigman, 2012). During scanning of complex arithmetic expressions, adults' gaze trajectories quickly move to the sub-expressions deepest in the syntactic hierarchy and thereafter proceed upwards along the syntactic tree (Schneider et al., 2012), suggesting that syntactic structure is extracted quickly and automatically. Further, after viewing complex algebraic expressions, sub-strings that constituted syntactic phrases within the expressions are recalled more easily than sub-strings that did not constitute syntactic phrases (Jansen et al., 2003). Apparently, syntactic structure influences encoding and subsequent recall of algebraic expressions. Several computational models assume that human processing of algebraic expressions begins with, and is subsequently guided by, syntactic parsing (Anderson, 2005, 2009; Jansen, Marriott, & Yelland, 2007).

1.1.2. Perceptual grouping

In perceptual grouping, as in syntax-based processing, evaluation of complex expressions begins with identification of simpler sub-expressions, but perceptual constraints rather than formal rules determine which symbols are grouped together to form sub-expressions (Landy, Allen, & Zednik, 2014). There is strong evidence that at least one such constraint – a tendency to group together symbols that are physically close to each other, consistent with the Gestalt principle of proximity (Wertheimer, 1938) – does indeed influence processing of symbolic expressions in arithmetic and algebra (Jiang, Cooper, & Alibali, 2014; Kirshner, 1989; Landy & Goldstone, 2007b, 2010). For example, violations of operator precedence rules are more common with expressions in which the operands surrounding a lower-precedence operator are more narrowly spaced than those surrounding a higher-precedence operator, as in "2+7 × 5" (Landy & Goldstone, 2010). Apparently, the perceptual constraint that closely-spaced symbols are more likely to be perceived as groups can sometimes override the formal rules that determine syntactic structure.

Importantly, while perceptual constraints may cause violations of formal rules, perceptual processing does not in general preclude formally correct performance. The reason is that perceptual constraints are flexible, and may evolve over time to come into closer alignment with such formal rules (Goldstone, Landy, & Brunel, 2011; Goldstone, Landy, & Son, 2010). For example, there is some evidence that adults experienced with arithmetic perceive higher-precedence operator symbols (e.g. ×, ÷) as more visually salient than lower precedence ones (e.g. +, –; Landy, Jones, & Goldstone, 2008). These differences in salience could lead to preferential grouping of operand symbols surrounding higher-precedence operators, resulting in formally correct order of evaluation. More generally, practice with symbol systems could lead to the development of automatic perceptual routines that effectively implement syntactic rules, without representing such rules explicitly. Consistent with this possibility, in a recent neuroimaging study, participants viewing arithmetic expressions of varying syntactic complexity showed effects of syntactic complexity on BOLD response in brain areas relating to early visual processing, while such effects were not found in areas associated with language (Maruyama, Pallier, Jobert, Sigman, & Dehaene, 2012; see also Friedrich & Friederici, 2009; Monti, Parsons, & Osherson, 2012; but see Scheepers et al., 2011).

1.1.3. Opportunistic selection

Opportunistic selection refers to prioritizing for evaluation sub-expressions which are relatively easy to evaluate. For example, in evaluating "25 + 13 – 3" one might begin by evaluating the subtraction ("13 – 3") because it is easier to evaluate than the addition ("25 + 13"). Opportunistic selection yields a formally correct answer in this case (13 – 3 = 10, 25 + 10 = 35), but not in all cases. In the similar problem "25 – 13 – 3," evaluating "13 – 3" first violates the rule of left-to-right order of operations and so yields an error.

There is some evidence that opportunistic selection does occur, and can even override formal rules of arithmetic. Linchevski and Livneh (1999; Herscovics & Linchevski, 1994) found that students frequently commit errors like that just mentioned, justifying their procedures by appeals to convenience (e.g. "when you do [operation] first, it becomes much easier"). However, these findings are not entirely conclusive for present purposes because the errors in question may have resulted from random slips, with convenience mentioned only as a post hoc rationalization. The present study addressed this possibility by comparing rates of order-of-operations errors between similar problems in which the (formally) low-priority sub-expressions either were or were not particularly easy to evaluate. Higher error rates for the former type of problem would provide strong evidence that opportunistic selection does occur and can override formal syntactic rules.

An important difference between opportunistic selection and syntactic parsing relates to the types of information to which they are sensitive. Because the ease of evaluating sub-expressions depends on the specific numbers involved, opportunistic selection is necessarily sensitive to the values of these numbers. Syntactic parsing, by contrast, depends only on syntactic structure, not on content, and should therefore be insensitive to the number values involved in an expression. This insensitivity to number values is

implicit in the computational models of algebra processing mentioned earlier[1] (Anderson, 2005, 2009; Jansen et al., 2007) and explicit in a recent model of arithmetic processing (Maruyama et al., 2012). Maruyama et al. (2012) proposed that such processing begins with a syntactic stage, in which syntactic structure is determined based on the positions and identities of bracket and operator symbols. Number symbols are only processed in the second, semantic stage, during which sub-expressions are evaluated in an order consistent with the extracted structure. The model is supported by the results of Schneider et al. (2012), which suggest that syntactic structure is already available to determine the sequence in which gaze fixations proceed through arithmetic expressions. Additionally, Maruyama et al. (2012) found that people more easily detect changes to operators when these were embedded in syntactically valid sub-expressions rather than invalid substrings, while detection of changes to number symbols showed no such influence, suggesting that operators are incorporated into the perceived structure more quickly than are numbers.

Despite this evidence, people do not always separate syntax and semantics as cleanly as Maruyama et al.'s (2012) proposal implies. During construction of algebraic and arithmetic expressions to represent situations, the semantic content of the situations influences the syntactic structure of the created expressions (Bassok, Chase, & Martin, 1998; Bassok, Wu, & Olseth, 1995; Fisher, Borchert, & Bassok, 2011; Martin & Bassok, 2005). The values of the numbers involved are among the semantic cues that exert such an influence (Bell, Fischbein, & Greer, 1984; Bell, Swan, & Taylor, 1981; Brissiaud & Sander, 2010; Fischbein, Deri, Nello, & Marino, 1985). For example, Bell et al. (1984) asked 12 and 13 year old students to write arithmetic calculations that would yield the answers to story problems. For problems in which the correct operation was multiplication, students usually wrote correct calculations when both operands were larger than 1 (e.g. "9×1.13") but usually chose the wrong operation when one operand was smaller than 1 (e.g. choosing "$2 - 0.14$" instead of "2×0.14"). These findings suggest the possibility that specific number values might also influence evaluation of existing arithmetic expressions, contrary to purely syntactic parsing but consistent with opportunistic selection. This possibility was tested in the present study.

1.2. Development

Given the evidence reviewed above, it is likely that people rely on multiple mechanisms, including syntactic parsing, perceptual grouping, and opportunistic selection, to evaluate complex arithmetic expressions. However, the relative importance of these mechanisms may change over time and with experience. Here we describe two possible developmental trajectories: a formal shift, in which reliance on syntactic parsing increases over development while reliance on perceptual grouping and opportunistic selection decreases, and a non-formal shift, in which reliance on the latter two mechanisms increases over development.

The formal shift view springs from the intuition that mature competence in arithmetic consists in mastery of the formal rules. In this case, older children with more experience in arithmetic should rely primarily on syntactic parsing, which consists in explicitly following these rules. Younger children might rely more on perceptual grouping, given that this mechanism does not require any knowledge of syntactic rules in order to operate. Similarly, younger children might engage in more opportunistic processing because they are not yet aware of, or skilled in using, the rules which constrain order of evaluation. The development of arithmetic competence would then involve a shift from greater reliance on perceptual grouping and opportunistic selection to greater reliance on syntactic parsing.

The reader will recognize this view as a specific version of the shift towards abstraction that has been proposed in many other domains (Chi et al., 1981; Chi & VanLehn, 2012; Gentner & Toupin, 1986; Gentner, 1988, 2003; Keil & Batterman, 1984; Keil, 1989; Piaget, 1952; Rattermann & Gentner, 1998; Vygotsky, 1962). Indeed, several researchers have proposed such a shift in the development of mathematical cognition in particular (De Lima & Tall, 2008; Novick, 1988; Tall, 1995, 2008). In this shift, mechanisms such as perceptual grouping and opportunistic selection serve as scaffolding that facilitates initial acquisition of procedural competence. This development paves the way for subsequent acquisition of formal knowledge via reification of procedural knowledge, which subsequently replaces the initial scaffolding (De Lima & Tall, 2008; Kirshner & Awtry, 2004; Sfard, 1991; Tall, 1995, 2008).

The non-formal shift view results from the intuition that mature competence consists not only in behaving in accordance with the formal rules, but also in doing so quickly and effortlessly. Explicit awareness of the rules of syntax during the course of problem-solving might actually interfere with fluency. On the other hand, perceptual grouping could promote fluency by enabling fast, effortless apprehension of the internal structure of arithmetic expressions. Consistent with this view, Goldstone and colleagues have argued that perceptual processes are fundamental to effective symbolic reasoning in mathematics and science (Goldstone, Landy, & Son, 2008; see also Goldstone et al., 2010; Kellman, Massey, & Son, 2010; Kellman & Massey, 2013; Landy et al., 2014). Similarly, opportunistic selection could promote fluency by enabling selection of efficient solution paths. Consistent with this view, procedural flexibility – the ability to solve problems using various methods rather than rigidly following a standard algorithm in all cases – is considered to be a hallmark of advanced mathematical skill (Rittle-Johnson & Star, 2009; Star, 2005). If this view is correct, then perceptual grouping and opportunistic selection are not merely scaffolds to be used on the path to competence and then cast aside, but are instead intrinsic to mature competence, and reliance on these mechanisms might actually increase with time and experience.

The non-formal shift view resembles shifts towards concreteness that have been proposed in other domains (Simons & Keil, 1995; Varma & Schwartz, 2011). For example, Varma and Schwartz (2011) argued that when comparing integer magnitudes, younger children rely on rules (e.g. positive numbers are larger than negative numbers), while older children rely in part on perceptual comparisons on a mental number line. In the present context, increasing reliance on perceptual grouping and opportunistic selection may be viewed as a shift towards concreteness because these mechanisms are influenced by relatively concrete information. Specifically, perceptual grouping relies on perceptual features such as symbol spacing, while syntactic structure, on which syntactic parsing relies, is an abstraction over perceptual features. Similarly, opportunistic selection relies on specific number values, while syntactic structure is an abstraction over specific number values. A common thread in these proposals is that reliance on perceptually or semantically concrete information increases with development.

In summary, changes with age and experience in the degree to which spacing between symbols and specific number values influence the evaluation of complex arithmetic expressions are diagnostic regarding the two conflicting views just described. The non-formal shift view predicts an increase in the influence of these formally irrelevant factors, while the formal shift view predicts the opposite. The main goal of the present study was to test these predictions.

[1] In Anderson's (2005, 2009) model, processing is partially dependent on identities of number symbols, because special production rules apply to expressions involving multiplication by 1 or addition of 0. However, these differences come into play only after expressions are parsed according to their syntactic structure.

1.3. Math garden

The study was conducted through a commercial website dedicated to mathematics education, Math Garden[2] (http://www.mathsgarden.com; Klinkenberg, Straatemeier, & van der Maas, 2011). Math Garden offers its users a platform for computerized adaptive practice (CAP) in mathematics. Math Garden has also been used extensively as a platform for psychological research (Jansen & Louwerse, 2013; Jansen et al., 2014; Jansen, de Lange, & van der Molen, 2013; Van der Ven, van der Maas, Straatemeier, & Jansen, 2013). In the latter capacity, Math Garden offered important advantages for the present study. First, its large and varied user base enabled us to achieve greater external validity than typically possible through collaboration with individual schools. Rather than being limited to a small sample of schools that agreed to participate, the sample was drawn from approximately 1000 participating schools, representing over 10% of all primary schools in the Netherlands. Second, Math Garden users practice on the website as assigned homework or practice for their regular mathematics classes, rather than to satisfy the demands of an experimenter in a laboratory. This realistic setting afforded greater ecological validity than possible in many laboratory experiments. In particular, we can have confidence that our results are not caused by idiosyncratic aspects of a laboratory environment.

A third advantage of Math Garden relates to its method of selecting practice problems for its users. Math Garden calculates and constantly updates ratings of the ability of its users and the difficulty of each individual problem in its bank, and uses these ratings as a basis for presenting each user with problems that are neither too easy nor too challenging. The algorithm by which this is accomplished is outlined here, and described in more detail in Appendix A and by Klinkenberg et al. (2011) and Maris and van der Maas (2012). Roughly, when a user answers a problem correctly, Math Garden increases the user's ability rating and decreases the problem's difficulty rating, while the opposite occurs when a problem is answered incorrectly. This algorithm is based on the Elo system for rating chess players' skills (Elo, 1978), with the user filling the role of one player, the problem playing the role of the other player, and correct (incorrect) solution corresponding to victory by the player (problem). Ratings shift more when the "winner" (user or problem) was rated much lower than the loser, and less when the winner was rated higher than the loser. The algorithm makes use of response time as well as accuracy information, so that problems solved accurately but slowly are rated as more difficult than problems solved accurately and quickly. The user ability and problem difficulty ratings are used to estimate the probability with which a given user would answer each problem correctly. When it is time for a user to receive a new problem, Math Garden selects a problem for which the user is estimated to have approximately a 75% chance of answering correctly. This approach avoids presenting users with problems that are too easy, which could cause boredom, or with problems that are too hard, which could cause discouragement and reduce motivation.

The advantage of the above algorithm for the present study is that the algorithm enables accurate assessment of problem difficulty in a realistic, motivating study environment. The central research questions of the study were addressed in part by analyzing the relative difficulties of different types of problems, as explained in the Method section. In a typical experiment, these relative difficulties might be assessed by having each participant solve all of the problems, or a subset of those problems selected either randomly or according to experimental condition. These approaches would have the drawback that many users could become bored or discouraged due to receiving problems that were either too easy or too hard for them. Instead, the problems were inserted into Math Garden's problem bank, and were encountered by users in the course of their regular use of the site. The Math Garden algorithm ensured that users would not receive problems that were too easy or too hard, thus promoting a relatively high level of engagement and motivation, as well as ensuring that users' answers would be highly informative as to their specific proficiency. At the same time, even though different users received different sets of problems that were *not* randomly selected, the relative difficulties of the problems can still be accurately assessed through analysis of the difficulty ratings generated by the algorithm.

2. Method

To investigate the mechanisms underlying evaluation of complex arithmetic expressions, problems were designed that could be evaluated in exactly two possible orders – one correct and one incorrect. Evaluation in the incorrect order always yielded an incorrect response, referred to as a "foil error." For example, with respect to the expression "$2 + 7 \times 5$," incorrectly adding 2 and 7 first to obtain 9, then multiplying 9 by 5, would yield the foil error 45. While holding syntactic structure constant, spacing between symbols and specific number values were manipulated in such a way that perceptual grouping and opportunistic selection, respectively, would either encourage or discourage correct order of evaluation. These manipulations were expected to affect the difficulty of correct evaluation and the frequency of foil errors. Further, if the formal shift view is correct, the effects of these manipulations should decrease with participants' age, while if the non-formal shift view is correct, they should increase with age.

2.1. Participants

Data were collected from 65,856 unique Math Garden users over a period of 23 months. Our analyses focus on data from users in Grades 4–8, equivalent to Grades 2–6 (approximately aged 8–12) in the USA. This range was chosen because students in these grade levels in the Netherlands have been exposed to all four basic arithmetic operations (+, −, ×, ÷), but not yet to algebra, in school. Users in these grade levels accounted for 58,660 (89.1%) of the 65,856 users in the total sample. Individual users contributed variable numbers of responses, ranging from 1 to 1902 (mean: 23.2, standard deviation: 36.8, median: 13). The total number of responses contributed by all users was 1,526,089, of which 1,357,092 (88.9%) came from users in Grades 4–8.[3] The number of responses contributed by users in each combination of sex and grade level are shown in Table 1.

As described below, the experiment involved 308 different arithmetic problems as stimuli, and our analysis of the data focused on items analyses rather than subject analyses. The number of responses received for individual problems ranged from 900 to 11,996 (mean: 4954.8, standard deviation: 3196.2, median: 3795). The number of responses received for each item from users in a single grade level ranged from 24 to 2824 (mean: 881.2, standard deviation: 656.7, median: 770.5).

[2] The third author is founder and member of the board of the company Oefenweb.nl that owns and manages Math Garden.

[3] In both the statistics reported here and the analyses of grade level reported subsequently, responses were classified by grade level according to the grade level of the user submitting the response at the time of submission. Thus, it was possible for a single user to provide data for multiple grade levels over the course of the study.

Table 1
Numbers of responses received from users in each combination of sex and grade level.

Sex	Dutch grade level (USA grade level)					Total
	4 (2)	5 (3)	6 (4)	7 (5)	8 (6)	
Male	114,287	160,737	177,098	171,403	135,618	759,143
Female	41,354	106,652	155,224	166,107	128,612	597,949
Total	155,641	267,389	332,322	337,510	264,230	1,357,092

2.2. Materials

198 arithmetic problems were constructed, divided among 4 problem families with between 24 and 60 problems each. These problems were inserted into Math Garden together with 1804 other arithmetic problems relating to another study not reported here. Examples of the problems relating to the present study are shown in Table 2. All of these problems were three-term arithmetic expressions, i.e. expressions involving three operands and two operators. No parentheses or brackets were used, so the correct order of execution of the two operations had to be determined based on the rules of arithmetic. For problems in Families 1 and 3, the relevant rule for determining correct order of evaluation was operator precedence: × must be executed before + and –. Problems in Families 2 and 4 involved two identical operators, either – or ÷. For these problems, the relevant rule for determining correct order of evaluation was left-to-right evaluation – the operator on the left must be executed first. The problems were designed so that executing operations in the incorrect order would always yield an incorrect response, termed a "foil error." Problems were selected so that correct responses and foil errors were always positive integers. Correct responses were selected so that correct responses and foil errors were always positive integers. Correct responses and foil errors for the examples in Table 2 are shown in the last two columns of the table.

The problems in each family were designed to investigate the effects of a particular property of arithmetic expressions on the difficulty of correct evaluation and likelihood of foil errors. These properties are listed in the column labeled "Property" in Table 2. Families 1 and 2 involved manipulations of spacing between symbols intended to detect usage of perceptual grouping mechanisms. Families 3–4 involved manipulations of ease of calculation of subexpressions via the specific number values involved, and were intended to detect opportunistic selection. The particular experimental manipulations employed were different for each family, and are shown in the column labeled "Factor 1" in the table. In each family, another property of secondary interest was also manipulated between problems; these properties are listed in the column labeled "Factor 2." The problems in each family were designed according to the factorial combinations of Factor 1 and Factor 2. Table 2 shows one example problem for each combination of factor levels within each family.

The problems in Family 1 each involved one addition and one multiplication, and foil errors were violations of operator precedence, in which addition was executed before multiplication (e.g. "2+7 ×5 = 45" by evaluating 2+7 first). Problems were constructed in sets of four, with all problems in a set sharing the same set of operands, the same correct response, and the same foil error. These four problems represented the factorial combinations of two factors: spacing and operator order. Spacing determined whether the operands would be narrowly spaced surrounding plus and normally spaced surrounding times (e.g. "2+7 ×5"), or vice versa (e.g. "2 + 7×5"), while operator order determined whether plus was the first (e.g. "2+7 ×5") or the second operator (e.g. "5 × 7+2"). Narrow spacing around plus, rather than times, was expected to increase difficulty and likelihood of foil errors. 15 sets of four problems each were generated, for a total of 60 problems in the family.

Problems in Family 2 each involved two subtractions or two divisions, and foil errors were violations of left-to-right order of execution, in which the second operation was executed before the first (e.g. "23 – 13–8 = 18" by evaluating 13–8 first). Problems were constructed in sets of two, with all problems in a set sharing the same operations (subtractions or divisions), the same set of operands, the same correct response, and the same foil error. The two problems differed with respect to spacing, i.e. in one problem, the first operator was surrounded by narrow and the second by normal spacing (e.g. "23–13 – 8"), and vice versa for the other problem (e.g. "23 – 13–8"). Narrow spacing around the second operator was expected to increase difficulty and the likelihood of foil errors. 15 sets of two problems each were generated for each of the two operations, subtraction and division, for a total of 60 problems.

Problems in Families 3 and 4 were designed to investigate whether expressions that were simple to calculate by virtue of the numbers involved would be prioritized for evaluation. In Family 3, the first operation was always subtraction and the second multiplication, and foil errors were violations of operator precedence, in which subtraction was executed before multiplication (e.g. "33 – 13 × 2 = 40" by evaluating 33 – 13 first). Problems were constructed in sets of two, with the subtraction operation simple to calculate in one problem and neutral to calculate in the other. In the simple case, the first and second operands shared a units digit, so that the units digit of their difference was evidently zero (e.g. "33 – 13 × 2"), while in the neutral case, the first and second operands did not share a units digit (e.g. "30 – 13 × 2"). Difficulty and foil error rates were expected to be greater when the subtraction was simple than when it was neutral. The problems in one set shared the same second and third operands, and differed only with respect to the first operand. Problem sets differed from each other with respect to the size of the difference between the first two operands, which could be either approximately 20 (e.g. "33 – 13 × 2") or approximately 100 (e.g. "115 – 15 × 3"). The differences were *exactly* 20 or 100 for the simple subtraction problems, and near to 20 or 100 for the neutral subtraction problems. Four sets of two problems each were created for differences of approximately 20, and 8 sets of two for differences of approximately 100, for a total of 24 problems.

Problems in Family 4 involved two subtractions or two divisions, and foil errors were violations of left-to-right order of execution, in which the second operation was executed before the first (e.g. "25 – 13 – 3 = 15" by calculating 13 – 3 first). Problems were constructed in sets of two, differing according to whether the first operation or the second operation was simple to calculate, if executed first. For problems involving subtraction, simplicity of calculation was manipulated as in Family 3, i.e. to make the first subtraction easy, the first and second operands would share a units digit (e.g. "32 – 12 – 6"), while to make the second subtraction easy, the second and third operands would share a units digit (e.g. "25 – 13 – 3"). For problems involving division, simplicity of calculation was manipulated by making either the quotient of the first and second operands equal to 10 (simple first operation, e.g. "120 ÷ 12 ÷ 2") or the quotient of the second and third operands equal to 10 (simple second operation, e.g. "2000 ÷ 50 ÷ 5"). Within each set, the three operands in one problem had similar though not necessarily identical values to those in the other problem. Difficulty and foil error rates were expected to be greater when the second, rather than the first, operation was simple to calculate. 15 sets of subtraction problems and 12 sets of division problems were created, for a total of 54 problems.

Table 2
Problem families, with examples.[a]

Family	Relevant rule	Property	Factor 1	Factor 2	Example	Correct response	Foil error
Family 1	Operator precedence	Spacing between symbols	Narrowly spaced plus	Plus first	2+7 × 5	37	45
			Narrowly spaced times	Plus first	2 + 7×5	37	45
			Narrowly spaced plus	Plus second	5 × 7+2	37	45
			Narrowly spaced times	Plus second	5×7 + 2	37	45
Family 2	Left-to-right evaluation	Spacing between symbols	Narrowly spaced 1st operator	Subtraction	23−13 − 8	2	18
			Narrowly spaced 2nd operator	Subtraction	23 − 13−8	2	18
			Narrowly spaced 1st operator	Division	64÷8 ÷ 4	2	32
			Narrowly spaced 2nd operator	Division	64 ÷ 8÷4	2	32
Family 3	Operator precedence	Ease of calculation	Simple subtraction	Difference ≈ 20	33 − 13 × 2	7	40
			Neutral subtraction	Difference ≈ 20	30 − 13 × 2	4	34
			Simple subtraction	Difference ≈ 100	115 − 15 × 3	70	300
			Neutral subtraction	Difference ≈ 100	112 − 15 × 3	67	291
Family 4	Left-to-right evaluation	Ease of calculation	Simple 1st operation	Subtraction	32 − 12 − 6	14	26
			Simple 2nd operation	Subtraction	25 − 13 − 3	9	15
			Simple 1st operation	Division	120 ÷ 12 ÷ 2	5	20
			Simple 2nd operation	Division	2000 ÷ 50 ÷ 5	8	200

[a] Following Dutch notational conventions, division was represented by a colon (":") rather than by the obelus ("÷") for all problems in the experiment. In the examples presented in this table and in the text of this article, the obelus is employed on the assumption that it is more familiar to most readers.

2.3. Procedure

Rather than recruiting participants specifically for this study, data were collected from Math Garden users in the course of their regular use of the Math Garden website. Users accessed the Math Garden system via web browsers from any location and at any time. Upon logging in, they could select any of a number of different "games" to play, each involving practice with a different type of mathematics problem. The problems relating to the present study were contained within a game called "Arithmetic Sequence." Only users who had previously demonstrated competence with basic arithmetic operations could access this game, while such users could access it as often as they wished. Each time a user accessed this game, they were presented with a sequence of problems for which the estimated probability of a correct response was near 75%, given the system's current ratings of the user's skill level and the difficulties of all problems in the system. These ratings were calculated using the Math Garden algorithm, as described in Appendix A.

The problems were presented one at a time. Fig. 1 illustrates the user interface for a single problem. Each problem appeared in a box at the top of the screen. Below the problem was a section for user input. Users employed a virtual keypad to enter a response, which would appear in the blue area, then clicked on "OK" to submit the response. If the response was correct, the answer box turned green and the next problem appeared automatically after 2 s. If the response was incorrect, the answer box turned red, the correct answer was displayed, and the next problem appeared after 10 s or after the user clicked a "continue" button, whichever came first.

Users were encouraged to respond both quickly and accurately by Math Garden's scoring system. (As described in the Appendix A, both accuracy and response time are taken into account by Math Garden's algorithm.) A row of coins was displayed at the bottom of the problem interface (Fig. 1). Users earned coins by solving problems quickly and accurately, and could later exchange these coins for virtual rewards. 20 coins appeared initially for each new problem, and the coins disappeared at a rate of one per second until a response was submitted. When a user submitted a response, they earned the number of coins currently displayed if the response was correct, and lost that number if the response was incorrect. If no response was submitted before all of the coins disappeared (i.e. after 20 s), the correct answer was automatically displayed and the user neither earned nor lost coins. This reward system was shared among all games in Math Garden and was therefore familiar to users.

Problems were presented in blocks of 15. However, users were free to quit at any time, including in the middle of a block. Users could also do as many blocks as they wished, and after stopping, could later return to the game as often as they wished. Thus, the number of problems completed, and the selection and sequence of these problems, was variable from user to user.

2.4. Measures

Items analyses were conducted to test the predictions described in the Materials. The reason for analyzing the data by items rather than by participants was that each participant solved a different, non-random subset of the entire set of problems. We expected our manipulations of problem features in each family to affect the difficulty of correct evaluation and likelihood of foil errors. To test these predictions, we calculated two measures for each problem: difficulty rating and foil error rate.

Difficulty ratings were calculated using the algorithm described in Appendix A. In brief, when a given student completed a given problem, the percent of coins gained or lost was used as a measure of performance. This measure takes both accuracy and response time into account, because fewer coins could be gained (or lost) following slower responses. Performance was predicted based on a formal model with two key parameters: problem difficulty rating and student ability level, and after each problem solving attempt, both of these parameters were updated based on actual performance. The difficulty ratings used for analysis were the most recently-updated values at the end of data collection, and therefore constitute aggregate measures of difficulty across all students who attempted each problem. The algorithm by which the ratings were updated is described in detail in Appendix A, while evidence for reliability of the final difficulty ratings is given in Appendix B.

Foil error rate was defined as number of foil errors divided by total number of responses for each problem. Foil error rates were calculated using only data from participants in the grade levels of interest, i.e. Dutch grades 4–8. Note that the calculation of difficulty ratings could not be constrained in the same way, because these ratings represented the sum total of all adjustments made by the algorithm throughout data collection. Similarly, foil error rates could be calculated separately for each grade level, while such calculations were not possible for difficulty ratings. Thus, analyses of effects of grade level were only performed with respect to foil error rates, not difficulty ratings.

Two secondary measures were also calculated for each problem: error rate (number of incorrect responses divided by number

Fig. 1. Screenshot illustrating the user interface for a single problem in Math Garden.

of responses) and proportion of foil errors (number of foil errors divided by number of incorrect responses). Difficulty rating was preferred to error rate as a measure of problem difficulty because the algorithm used to select problems would tend to diminish differences in error rates by preferentially assigning easier problems to less skilled users and more difficult problems to more skilled users. Foil error rate was preferred to proportion of foil errors as a measure of the likelihood of foil errors because our predictions pertained to absolute frequency of foil error responses. Analyses of error rate and proportion of foil errors typically yielded results consistent with those reported for difficulty rating and foil error rate, respectively. Exceptions are noted in the Results.

2.5. Analyses

Each family involved manipulation of a formally extraneous problem feature, termed factor 1 (Table 2), which was predicted to affect problem difficulty and likelihood of foil error responses. We tested these predictions by submitting difficulty ratings and foil error rates of the problems in the family to a one-way ANOVA with factor 1 as a repeated-measures factor. When this analysis yielded a significant result, we tested whether the effect on foil error rates changed with participants' age/experience by adding grade level as a repeated-measure numeric predictor to the analysis of foil error rates. Finally, when a significant interaction between factor 1 and grade level was found, linear regression was used to assess the magnitude of the interaction by regressing the difference in mean values between different levels of factor 1 against grade level. The secondary problem features, termed factor 2 (Table 2), did not pertain to our main research questions and thus are not included in the analyses reported below. Analyses including factor 2 in each family are reported in Supplementary Materials. In general, including factor 2 did not affect the main analysis findings. Exceptions are noted in the Results.

3. Results

3.1. Summary

The key results of our analyses for all four problem families are summarized in Table 3. Difficulty ratings and foil error rates were both significantly higher when narrow spacing was used around operators that should have been executed second, rather than first. This result held whether the formally correct order of operations was determined by rules of operator precedence (Family 1) or left-to-right evaluation (Family 2). In both cases, the magnitude of the spacing effect increased significantly with grade, by about 1.9% per grade level. Subtraction and division operations were more likely to be evaluated prematurely if doing so made them simpler to evaluate (Families 3–4). When correct order of operations was determined by operator precedence (Family 3), this result was reflected in the analysis of foil error rate, but not difficulty rating, while when correct order was determined by left-to-right evaluation (Family 4), the result was evident in the analyses of both foil error rates and difficulty ratings. In the latter case only, the magnitude of the effect of simplicity of calculation increased with grade, by about 1.2% per grade level. The detailed analyses of each problem family are presented below.

3.2. Family 1 (spacing between symbols – operator precedence)

For problems in Family 1 (e.g. "5 × 7+2"), narrow spacing around the plus sign was expected to increase difficulty and frequency of foil errors, in which addition is executed before multiplication. Repeated-measures ANOVA revealed significant effects of spacing on both difficulty ratings, $F(1, 14) = 36.03$, $p < .001$, $\eta_g^2 = .251$, and foil error rates, $F(1, 14) = 74.87$, $p < .001$, $\eta_g^2 = .511$.[4] As shown in Fig. 2, higher difficulty ratings and higher foil error rates resulted when narrow spacing surrounded the plus sign rather than the times sign. Thus, spacing between symbols had the predicted effect.

To test for a developmental trend in the magnitude of this effect, grade level was added to the analysis of foil error rate as a numeric repeated measure. The main effect of grade level was significant, $F(1, 14) = 77.77$, $p < .001$, $\eta_g^2 = .545$, indicating that foil error rates increased from the earlier to the later grades (e.g. Grade 4: 9.1%, Grade 8: 13.6%). This increase was accompanied by a decrease in response time (e.g. average response time was 9.26 s in Grade 4, but 8.13 s in Grade 8). However, the fact that overall error rates decreased concurrently (e.g. Grade 4: 34.6%, Grade 8: 31.3%) argues against the increase in foil errors resulting simply from carelessness or a speed-accuracy tradeoff. Similarly, the decrease in overall error rates makes it unlikely that the increase

[4] Problems which belonged to the same set and differed only with respect to the order of plus and times, such as 2+7×5 and 5×7+2, were treated as a single item for these and all other items analyses in Family 1. All reported effects remained significant if such pairs were treated as distinct items instead of as a single item.

Table 3
Summary of items analyses. Factor 1 indicates the primary factor of interest for each problem family. "Main effect," main effect of Factor 1 in ANOVA. "Interaction with grade," interaction of Factor 1 with grade level in ANOVA. "Regression against grade," linear regression against grade level of the difference in means at different levels of Factor 1. Significant p values are marked as $^{*}(p < .05)$, $^{**}(p < .01)$, or $^{***}(p < .001)$.

Family	Relevant rule	Factor 1	Measure	Main effect		Interaction with grade		Regression against grade	
				p	η_g^2	p	η_g^2	p	β
1	Operator precedence	Narrowly spaced plus or times	Difficulty	<.001***	.251	–	–	–	–
			Foil rate	<.001***	.511	<.001***	.433	<.001***	0.019
2	Left-to-right evaluation	Narrowly spaced 1st or 2nd operator	Difficulty	<.001***	.148	–	–	–	–
			Foil rate	<.001***	.424	<.001***	.182	<.001***	0.019
3	Operator precedence	Simple or neutral subtraction	Difficulty	.085	.092	–	–	–	–
			Foil rate	.002**	.272	.562	.003	–	–
4	Left-to-right evaluation	Simple 1st or 2nd operation	Difficulty	<.001***	.489	–	–	–	–
			Foil rate	<.001***	.434	<.001***	.047	.013*	0.012

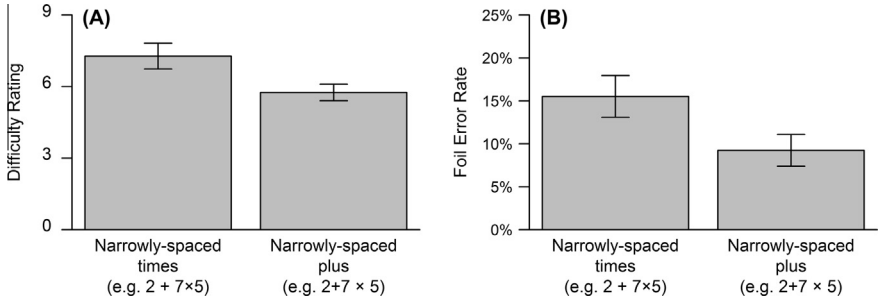

Fig. 2. (A) Difficulty ratings and (B) foil error rates for Family 1 problems.

in foil errors was caused by comparable problems being assigned to less able students at higher grade levels.

Critically, the interaction of grade level with spacing was significant, $F(1, 14) = 73.62$, $p < .001$, $\eta_g^2 = .433$. As shown in Fig. 3, the effect of spacing increased with grade level. To assess the rate of this increase, the mean difference in foil error rates between problems in which plus or times was narrowly spaced was submitted to linear regression with grade level as a predictor. The regression was significant, $F(1, 73) = 58.5$, $p < .001$, $\beta = 0.019$, indicating that the magnitude of the spacing effect increased by about 1.9% for each increase of one grade level.

3.3. Family 2 (spacing between symbols – left-to-right evaluation)

For problems in Family 2 (e.g. "23 – 13–8"), narrow spacing around the second operator was expected to increase problem difficulty and frequency of foil errors, in which the second operation is executed before the first. Repeated-measures ANOVA found significant effects of spacing on both difficulty ratings, $F(1, 29) = 37.97$, $p < .001$, $\eta_g^2 = .148$, and foil error rates, $F(1, 14) = 40.81$, $p < .001$, $\eta_g^2 = .424$. As shown in Fig. 4, higher difficulty ratings and higher foil error rates resulted when the second rather than the first operator was narrowly spaced. Again, spacing between symbols had the predicted effect.

Grade level was next added to the analysis of foil error rate as a repeated measure. The main effect of grade level was significant, $F(1, 29) = 95.66$, $p < .001$, $\eta_g^2 = .173$, indicating an increase in foil error rates from the earlier to the later grades (e.g. Grade 4: 6.5%, Grade 8: 12.0%). This increase was accompanied by a decrease in overall error rates (e.g. Grade 4: 45.3%, Grade 8: 36.7%) and no change in average response times (e.g. Grade 4: 9.08 s, Grade 8: 9.00 s).

The critical interaction of grade level with spacing was significant, $F(1, 29) = 27.01$, $p < .001$, $\eta_g^2 = .082$. As shown in Fig. 5, the effect of spacing increased with grade level. (Addition to the analysis of the factor of secondary interest in this family, i.e. whether the operation was subtraction or division, revealed a significant 3-way interaction of spacing and grade level with operation, $p < .001$. The 2-way interaction of spacing and grade level was primarily driven by division problems and was not present for

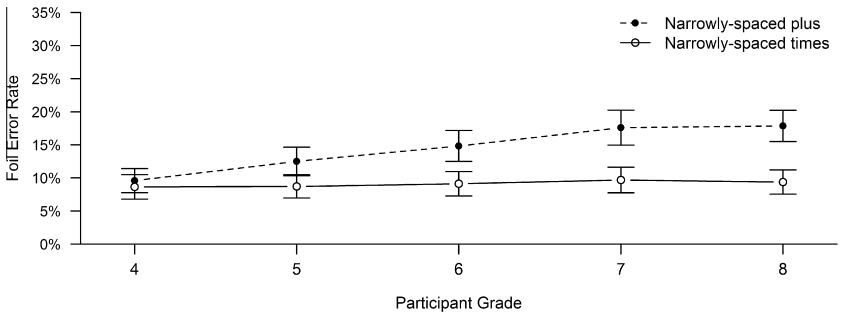

Fig. 3. Foil error rates for Family 1 problems, by spacing and Dutch grade level.

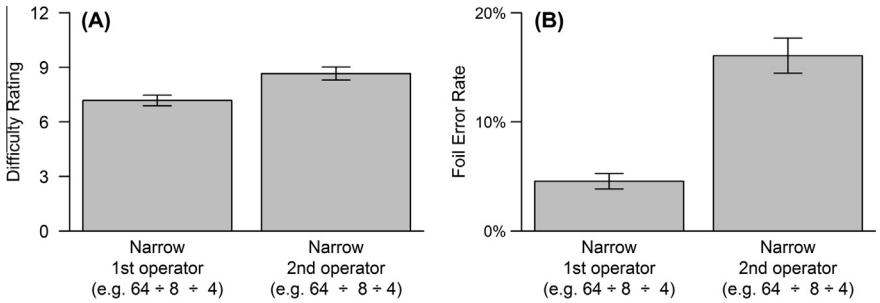

Fig. 4. (A) Difficulty ratings and (B) foil error rates for Family 2 problems.

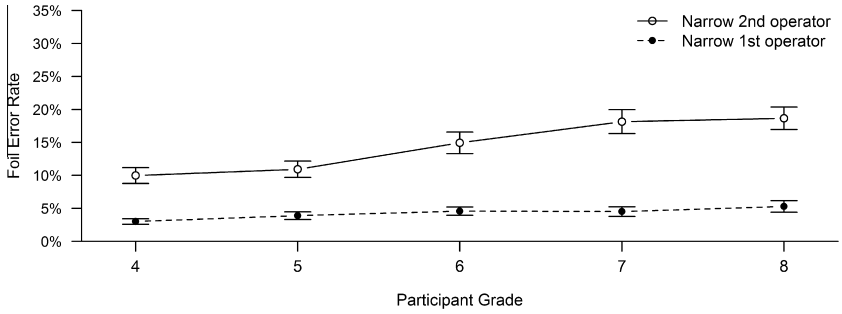

Fig. 5. Foil error rates for Family 2 problems, by spacing and Dutch grade level.

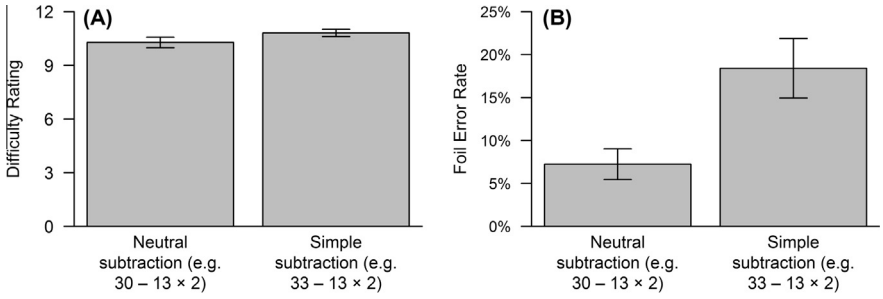

Fig. 6. (A) Difficulty ratings and (B) foil error rates for Family 3 problems.

subtraction problems. Further detail is provided in Supplementary Materials.) The mean difference in foil error rates between problems in which the second or first operator was narrowly spaced was submitted to linear regression with grade level as a predictor. The regression was significant, $F(1, 148) = 13.2$, $p < .001$, $\beta = 0.019$. Thus, the magnitude of the spacing effect increased by about 1.9% for each increase of one grade level.

3.4. Family 3 (ease of calculation – operator precedence)

For problems in Family 3 (e.g. "33 – 13 × 2"), simplicity of calculation of the subtraction operation was expected to increase problem difficulty and frequency of foil errors, in which subtraction is evaluated before multiplication. Repeated-measures ANOVA found only a marginally significant effect of simplicity of subtraction on difficulty ratings, $F(1, 11) = 3.57$, $p = .085$, $\eta_g^2 = .092$, but a significant effect on foil error rates, $F(1, 11) = 16.35$, $p = .002$, $\eta_g^2 = .272$. (When approximate size of difference in the subtraction operation, either 20 or 100, was included as a factor in the analysis, the effect of simplicity of subtraction on difficulty ratings became significant, $p = .027$.) As shown in Fig. 6B, foil error rates were higher when the subtraction operation was simple to execute than when it was neutral; difficulty ratings tended to show the same trend, though not as strongly (Fig. 6A). (Similarly, the effect of simplicity of subtraction on proportion of foil errors was significant, $p < .001$, while the effect on error rates was not significant, $p = .757$.)

When grade level was added to the analysis of foil error rate, the main effect of grade level was significant, $F(1, 11) = 19.28$, $p = .001$, $\eta_g^2 = .118$, indicating an increase in foil error rates from the earlier to the later grades (e.g. Grade 4: 7.7%, Grade 8: 13.4%). In contrast to Families 1–2, this increase was accompanied by an increase in overall error rates (e.g. Grade 4: 33.5%, Grade 8: 37.3%), although this increase was smaller than the increase in foil error rates, implying a slight decrease in frequency of non-foil errors. (Response times also increased with age in this family, e.g. Grade 4: 7.34 s, Grade 8: 8.72 s.) However, the interaction of simplicity of subtraction with grade level did not reach significance, $F(1, 11) = 0.358$, $p = .562$. Thus, we did not perform a post hoc regression analysis as in other families to assess the size of the grade effect.

3.5. Family 4 (ease of calculation – left-to-right evaluation)

For problems in Family 4 (e.g. "25 – 13 – 3"), problem difficulty and frequency of foil errors, in which the second operation is executed prematurely, were expected to increase when the second operation was simple to calculate. Repeated-measures ANOVA found significant effects of which operation was simple to calculate

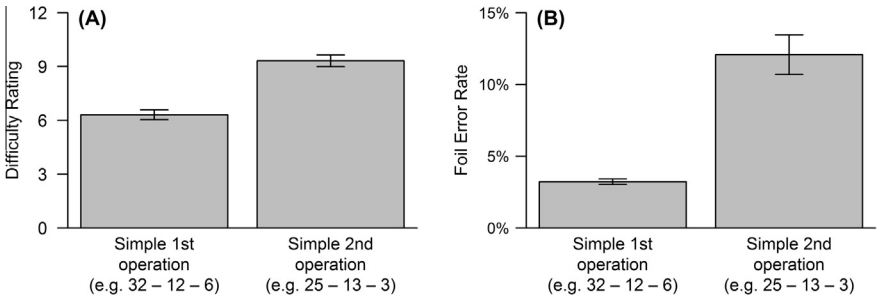

Fig. 7. (A) Difficulty ratings and (B) foil error rates for Family 4 problems.

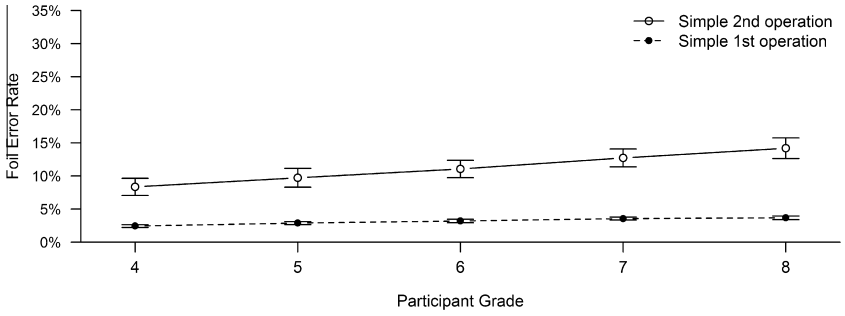

Fig. 8. Foil error rates for Family 4 problems, by ease of calculation and Dutch grade level.

on both difficulty ratings, $F(1, 26) = 67.72$, $p < .001$, $\eta_g^2 = .489$, and foil error rates, $F(1, 26) = 36.51$, $p < .001$, $\eta_g^2 = .435$. As shown in Fig. 7, difficulty ratings and foil error rates were higher when the second operation was simple to calculate. Thus, simplicity of calculation had the predicted effect.

Addition of grade level as a repeated measure to the analysis of foil error rates revealed a significant main effect of grade level, $F(1, 26) = 97.70$, $p < .001$, $\eta_g^2 = .106$, indicating an increase in foil error rates from the earlier to the later grades (e.g. Grade 4: 5.4%, Grade 8: 8.9%), despite a decrease in overall error rates (e.g. Grade 4: 45.1%, Grade 8: 34.7%) and no change in response times (e.g. Grade 4: 8.93 s, Grade 8: 8.82 s). The critical interaction of grade level with spacing was significant, $F(1, 26) = 44.88$, $p < .001$, $\eta_g^2 = .047$. As shown in Fig. 8, the effect of which operation was simple to calculate increased with grade level. The mean difference in foil error rates between problems in which the second or first operation was simple to calculate was submitted to linear regression with grade level as a predictor. The regression was significant, $F(1, 133) = 6.33$, $p = .013$, $\beta = 0.012$. Thus, the magnitude of the ease of calculation effect increased by about 1.2% for each increase of one grade level.

4. Discussion

Below, we briefly discuss the implications of the main effects of our experimental manipulations regarding the mechanisms underlying competence in arithmetic. We then discuss in detail the interactions of these effects with age, including their implications regarding the development of arithmetic competence, theories of mathematical development more generally, and educational practice in mathematics.

4.1. Mechanisms supporting arithmetic competence

4.1.1. Perceptual grouping

Consistent with the view that evaluation of complex arithmetic expressions relies in part on perceptual grouping mechanisms to identify and prioritize sub-expressions for evaluation, narrow spacing surrounding operand symbols in problem Families 1 and 2 increased the probability that those operands would be evaluated first, even when doing so constituted a violation of the rules for order of evaluation. This finding echoes those of several previous studies (Jiang et al., 2014; Kirshner, 1989; Landy & Goldstone, 2007b, 2010). This replication is encouraging, considering several differences between the present study and previous ones. First, while previous studies were conducted in the United States and Canada, the present study employed Dutch participants, suggesting that the effects of spacing do not likely result from idiosyncratic properties of particular educational systems. Second, the present study was embedded into participants' regular mathematics study under conditions of (presumably) relatively high motivation, suggesting that effects of spacing generalize beyond the laboratory settings employed in previous studies. Finally, the present study employed primary school students, in contrast to the previous studies mentioned above, all of which employed university students. The fact that similar effects were found in all of these studies despite their methodological differences suggests that these effects are both robust and general.

4.1.2. Opportunistic selection

In problem Families 3 and 4, sub-expressions were more likely to be evaluated prematurely, in violation of correct order of operations, when they were easy to calculate due to the specific number values involved. Corroborating earlier findings of Linchevski and Livneh (1999; Herscovics & Linchevski, 1994), this result is consistent with people prioritizing easy-to-calculate sub-expressions for evaluation, a mechanism we have dubbed "opportunistic selection." More generally, the result suggests that a complete theory of human evaluation of symbolic expressions in arithmetic and, likely, similar domains should account not only for syntactic and perceptual constraints, but also for procedural constraints such as a preference for easier actions over harder ones. Such procedural constraints are not reflected in several existing models of arithmetic and algebraic expression processing (Anderson, 2005, 2009; Jansen et al., 2007; Maruyama et al.,

2012). Accommodating such constraints within such models would likely require relaxing the assumption that parsing of structure precedes and influences evaluation, but is not influenced by it.

This conclusion dovetails with several studies (Alibali, Phillips, & Fischer, 2009; Crooks & Alibali, 2013; McNeil & Alibali, 2004) suggesting that knowledge of procedures can impact perception of arithmetic expressions. When asked to reconstruct equivalence problems, such as "3+4+6=3+__," after brief presentation, students often place the equals sign to the right of all operators, as in "3+4+6 +3=__." McNeil and Alibali (2004) argued that such errors result from assimilation into a familiar perceptual pattern – namely, operands on the left, answer on the right – associated with the well-practiced procedure of applying all given operations to all given operands. Consistent with this view, students who learned to solve the problems by equalizing the two sides of the equations were subsequently less likely to commit similar reconstruction errors, presumably because this procedure highlights the possibility of operations on both sides of the equation (Alibali et al., 2009). Thus, symbolic expressions are often perceived in a manner consistent with application of known procedures. Extending this conclusion, the present findings suggest that procedural knowledge can influence not only perception of physical features such as ordinal position of symbols but also perception of non-physical properties such as the internal structure of expressions, and that the ease with which known procedures may be executed is one factor that can exert such an influence.

A possible alternative explanation for the present findings is that the supposedly easy-to-evaluate sub-expressions in problem Families 3 and 4 were prioritized for evaluation due to perceptual grouping rather than opportunistic selection. The operands in these sub-expressions always shared one or more digits (e.g. "120" and "12" in "120 ÷ 12") and may, therefore, have been perceived as groups due to perceptual similarity of their operands, consistent with the Gestalt principle of similarity (Wertheimer, 1938). However, if perceptual similarity alone were responsible for the observed effects, one might expect these effects to be even stronger in expressions containing sub-expressions with identical rather than merely similar operands. In fact, the opposite result was found in two problem families not reported in the current study. The results from these families did reveal a tendency to prioritize evaluation of sub-expressions with identical operands (e.g. "4 + 4" in the expression "7 × 4 + 4"), but this effect appeared in only one of the two problem families and was much smaller than those observed in Families 3 and 4. Thus, perceptual grouping is unlikely to account completely for the results obtained from Families 3 and 4.

4.2. Development of arithmetic competence

The principal contribution of the present study is the finding that the effects of both symbol spacing and ease of calculation on order of evaluation increased with grade level. The fact that prioritization of closely-spaced sub-expressions increased with grade level suggests that reliance on perceptual grouping increases with age and arithmetic experience. Similarly, the developmental trend in prioritization of easy-to-evaluate sub-expressions suggests that opportunistic selection also increases with age and experience. Together, these results imply that the development of the ability to evaluate complex arithmetic expressions in the correct order cannot be fully characterized in terms of increasingly consistent and correct use of syntactic parsing, and, in fact, experience often leads to less strict adherence to the formal properties of mathematics.

Several aspects of the data permit elimination of alternative explanations of our results. First, because participants were not randomly assigned to problems within each grade level, it is possible that the participants assigned to a given problem at later grade levels tended to be less competent than those assigned to the same problem at earlier grade levels. If so, effects of increasing grade level might actually be effects of decreasing competence. However, in this case, not only foil error rates but also overall error rates should increase with grade level. In fact, the opposite occurred in the three problem families (1, 2, and 4) which showed significant interactions involving grade level. A second possibility is that the higher foil error rates reflect age-related increases in procedural flexibility rather than increased reliance on perceptual grouping and opportunistic selection (Rittle-Johnson & Star, 2009; Star, 2005). While procedural flexibility is typically associated with skill in mathematics, increases in flexibility could have the side effect of increasing foil error rates because foil errors result specifically from evaluating expressions in orders other than the standard order. However, while this possibility could explain why foil error rates increased with grade level concurrent with decreases in overall error rates, it cannot explain why the effects of symbol spacing and ease of calculation on foil error rates should also increase with grade level.

Why should reliance on perceptual grouping and opportunistic selection increase with age and experience? We suspect the reason to be that these mechanisms reduce the time and effort required to encode and manipulate symbolic expressions. Fluency – the ability to perceive and act quickly and with minimal effort – is considered essential to expertise in general (Chase & Simon, 1973; Koedinger, Corbett, & Perfetti, 2012) because fluency enables experts to deal with complex situations by reducing the mental resources expended on processing details. Further, perceptual mechanisms can support the development of fluency and expertise in mathematics in particular (Goldstone et al., 2008, 2010; Kellman & Massey, 2013; Koedinger & Anderson, 1990). It is plausible that opportunistic selection also contributes to fluency, and thus to expertise, because opportunistic selection by definition involves choosing solution paths that reduce subsequent effort.

Expertise certainly does not depend on fluency alone. Fluently-performed procedures may still be incorrect, as indeed illustrated by the present study. Further, a sense of fluency can inhibit analytical reasoning (Alter, Oppenheimer, Epley, & Eyre, 2007; Diemand-Yauman, Oppenheimer, & Vaughan, 2010; Oppenheimer, 2008) and cause overestimation of one's own understanding (Bjork, Dunlosky, & Kornell, 2013). Nevertheless, the present results suggest that increasing reliance on mechanisms that contribute to fluent performance may occur naturally in the development of mathematical competence. The implications of this conclusion for educational practice are considered in the next section.

While a tendency to adopt mechanisms that support greater fluency is one possible explanation for age-related increases in effects of symbol spacing and number values on order of evaluation, these increases could result from statistical learning processes unrelated to fluency *per se*. More specifically, in written arithmetic, adults tend to use narrower spacing around higher-precedence operator symbols (Landy & Goldstone, 2007a). Thus, even though symbol spacing is irrelevant to operator precedence, narrow spacing may be statistically associated with higher operator precedence in written arithmetic. Students may pick up on this association and thus tend to prioritize narrowly-spaced sub-expressions (e.g. Landy et al., 2008). Similarly, ease of calculation may be a valid cue for determining correct order of evaluation in students' experience, even though there is no formal reason this should be true. That is, expressions in which it is correct to evaluate the more easily-calculated sub-expression first, such as "25 – 15 – 3," may simply be more common than those in which it is incorrect to do so, such as "25 – 13 – 3." The present findings do not allow us to eliminate these possibilities (though, see Landy & Goldstone, 2010 for an argument that effects of symbol spacing

in algebra cannot be attributed entirely to learned statistical associations). However, even if increasing effects of formally irrelevant factors (i.e. symbol spacing and number values) on order of evaluation are a consequence of statistical learning rather than a drive towards fluency, the conclusion still stands that such a development is not adequately characterized in terms of increasing reliance on formal syntactic structure.

The developmental trends identified in the present study may not continue indefinitely. Children's performance might become increasingly aligned with formal rules at grade levels higher than those included in the study, leading to a decrease in the observed effects of symbol spacing and number values at that time. McNeil (2007) observed just such a non-monotonic trend in the development of children's understanding of mathematical equivalence over the ages 7–11. Accuracy in solving equivalence problems decreased between ages 7 and 9, presumably due to reinforcement through practice of a procedural understanding of the equals sign, but then increased between ages 9 and 11 as children gained a relational understanding of the equals sign.

With respect to the present findings it is important to note that a decrease, at higher grade levels, in the size of the effects observed would not necessarily reflect an increase in reliance on syntactic parsing. Instead, such a decrease might result from adjustment of mechanisms such as perceptual grouping and opportunistic selection to bring them into closer alignment with formal syntax (Goldstone et al., 2011, 2010). For example, experience with arithmetic and algebra might differentially increase the salience of higher-precedence operators, allowing attentional mechanisms more effectively to implement operator precedence rules (Landy et al., 2008; Landy, 2007). On the other hand, while the effects of spacing observed in Families 1 and 2 may decrease at higher grade levels, they are still present even among university students (Jiang et al., 2014; Kirshner, 1989; Landy & Goldstone, 2007b, 2010). Thus, alignment of perceptual grouping mechanisms with formal syntactic rules may not occur completely or for all learners, even after considerable formal instruction in mathematics. Future research should test whether the same point holds in the case of opportunistic selection, by attempting to replicate the present findings regarding effects of specific number values among more mathematically sophisticated individuals, such as undergraduate or graduate students in STEM (Science, Technology, Engineering, and Mathematics) departments.

4.3. Implications regarding mathematical development

In general, the present findings challenge the view that a shift towards abstraction (Chi et al., 1981; Chi & VanLehn, 2012; Gentner & Toupin, 1986; Gentner, 1988, 2003; Keil & Batterman, 1984; Keil, 1989; Piaget, 1952; Rattermann & Gentner, 1998; Vygotsky, 1962) can fully account for the development of mathematical cognition. In mathematics, the idea of a shift towards abstraction appears in the guise of a shift towards formal thinking, characterized by reliance on formal rules and axioms. Such a shift implies that formally irrelevant factors, such as those manipulated in the present study, should exert a decreasing influence on performance over time. As an example, Briars and Siegler (1984) found that very young children's conception of counting was heavily dependent on formally extraneous features of counting procedures, such as whether items were counted in standard left-to-right order. Older children were better able to distinguish definitional features of correct counting – that is, one-to-one correspondence between count words and counted objects – from formally extraneous features.

Tall's (1995, 2008; De Lima & Tall, 2008) "three worlds" framework describes a similar shift in broader mathematical cognitive development from childhood to adulthood. In this framework, mathematical thought belongs to embodied, symbolic, or formal worlds, or to blends among these. Development begins with the embodied and symbolic world, while the formal world later becomes primary, a change referred to as a "transition to formal thinking." Competence in symbolic manipulations is a necessary prerequisite for this transition and remains important after it, because formal thinking continues to rely on symbolic representations. The framework therefore acknowledges that such competence plays a foundational and persistent role in mathematical cognitive development. However, as we have argued, symbolic competence could result from a variety of different cognitive mechanisms, such as syntactic parsing, perceptual grouping, and opportunistic selection in the case of arithmetic evaluation. A progression towards formal thinking suggests increasing reliance on the mechanisms most compatible with such thinking, such as syntactic parsing, and decreasing reliance on mechanisms that emphasize formally extraneous factors, such as perceptual grouping and opportunistic selection. It certainly does not predict increasing reliance on the latter mechanisms even in cases when they conflict with formal rules. Yet, just such an increase was observed in the present study. While this finding is not actually contradictory with an eventual progression towards formal thinking, it does suggest that such a progression is not the whole story.

Do the present findings reflect developmental trends idiosyncratic to arithmetic, or do similar changes also appear in other mathematical domains? In fact, similar trends have been observed in research on the numeric cognitive development. Representation of numerical magnitudes is believed to rely on a visuo-spatial mechanism sometimes called a "mental number line" (Fischer & Shaki, 2014; McCrink & Opfer, 2014). One source of evidence is the Spatial-Numerical Association of Response Codes (SNARC) effect (Dehaene, Bossini, & Giraux, 1993), in which manual responses associated with small numbers are given more easily with the left than the right hand, while the reverse is true for large numbers, consistent with mapping of numeric magnitudes onto a (left–right) spatial axis. The automaticity and magnitude of the SNARC effect appears to increase with age (Van Galen & Reitsma, 2008; Wood, Willmes, Nuerk, & Fischer, 2008), suggesting that the visuo-spatial properties of the mental number line exert an increasing influence over development. Another source of evidence is distance effects, in which the speed or accuracy of comparison between numbers depends on the numeric distance between them (Moyer & Landauer, 1967), consistent with effects of distance on discriminability of locations in a spatial continuum. A recent study found distance effects in the comparison of positive and negative integers among adults, but not among 6th grade children (Varma & Schwartz, 2011). Varma and Schwartz (2011) concluded that children rely on a rule (i.e. any positive integer is greater than any negative one) for such comparisons, while adults rely additionally on visuo-spatial representations of positive and negative integers. In sum, findings on numeric cognition parallel the present results in suggesting that reliance on perceptual mechanisms may increase over the course of mathematical cognitive development.

4.4. Implications regarding mathematics education

From the perspective of educational practice in mathematics, two quite different attitudes are possible towards the observation that some formally extraneous factors exert an increasing influence on performance over time. On the one hand, one might view these trends as undesirable and potentially mutable, or at least remediable, through education. That is, even if mathematics students' perceptions of symbolic expressions are influenced by formally extraneous factors, they should not be so, and it is the job of educators to foster greater attention to, and understanding of,

formal properties such as syntactic structure. Kirshner and Awtry (2004) described a curricular approach consistent with this philosophy, termed "Lexical Support Systems" (LSS). LSS emphasizes the deliberate and systematic use of terminology designating structural components such as *terms* and *factors*, with the intention of providing "an explicit declarative account of the conventions for parsing algebraic expressions." While Kirshner and Awtry's (2004) proposal is distinctive in its emphasis on formal syntactic structure as an alternative to "mindless matching of visual patterns," many other researchers have associated automatic perceptual and procedural mechanisms with lack of conceptual understanding in mathematics (Kamii & Dominick, 1998; McNeil & Alibali, 2005; Richland, Stigler, & Holyoak, 2012). One might naturally infer from this association that reliance on such automatic mechanisms is itself a barrier to understanding, to be removed by appropriate instruction.

An alternative attitude, however, is that automatization of the processes of encoding and evaluating arithmetic expressions is desirable and should be encouraged (Goldstone et al., 2008, 2010; Kellman et al., 2010; Kellman & Massey, 2013). Furthermore, such automatization need not be viewed as an alternative to conceptual understanding. Instead, perceptual and procedural automatization can reduce the cognitive effort required to implement formal principles, thereby facilitating performance consistent with such principles. A few recent studies have tested curricular interventions designed to achieve such goals (Kellman et al., 2010; Ottmar, Landy, & Goldstone, 2012; Ottmar, Weitnauer, Landy, & Goldstone, 2015). Kellman et al. (2010), for example, applied perceptual learning principles to instruction in the relations between alternate representations (e.g. graphs, equations) of linear functions. Participants performed a representation matching task under time pressure, completing a large number of trials in a relatively short time. Strikingly, the test intervention led to better performance on a representation translation task, relative to a control condition that received direct practice on that task. As this study illustrates, interventions aimed at fostering development of automatic perceptual and procedural routines may place greater emphasis on repeated practice, perhaps under time pressure, while explicit instruction and discussion, an important component of Kirshner and Awtry's (2004) proposal, may play a smaller role in such interventions.

The results of the present study are in some respects consistent with both of the attitudes described above. On the one hand, the findings suggest that students may develop automatic perceptual and procedural routines over time even if not deliberately trained to do so. It may be impossible to prevent this process, but possible to guide it so that students acquire routines that more closely approximate formally correct procedures. On the other hand, the findings also suggest that pre-existing constraints of these perceptual and procedural mechanisms can be a systematic source of error. Emphasizing the importance of syntactic, rather than perceptual or procedural, constraints may be all the more important in this context. We suspect that this difference of opinion will be difficult to resolve entirely without a complete theoretical account, not only of the mechanisms underlying human processing of symbolic expressions, but also of the processes underlying the development of those mechanisms. It is hoped that the present findings will contribute to the eventual development of such an account.

A final implication for education regards spacing between symbols. Regardless of whether reliance on perceptual grouping is encouraged or discouraged, the fact that students do rely on it and that such reliance may even increase with age suggests that instructors should be cognizant of the potential influence of symbol spacing on students' learning and performance. Specifically, students may be confused by arithmetic expressions in which spacing is inconsistent with syntactic structure. On the one hand, instructors should take care not to produce such expressions when demonstrating for students. On the other hand, students themselves may inadvertently introduce spacing/structure inconsistency when writing arithmetic expressions and thereby create unnecessary obstacles to their own practice and learning. This problem may be particularly common or serious for students with motor difficulties. Instructors could potentially alleviate such problems by monitoring students' writing and correcting inconsistent spacing when it does occur.

Acknowledgements

This research was supported in part by National Science Foundation (United States) REESE Grant No. 0910218 and Department of Education (United States) IES Grant No. R305A1100060.

Appendix A

The algorithm underlying Math Garden maintains ratings of user ability θ_j and of problem difficulty β_i for each of the users and problems in its database. These ratings are updated each time a user attempts to solve a problem. The ratings are also used to select which problem will be presented next to a given user, with the goal that users should receive problems that are neither too easy nor too difficult.

Central to the calculation of θ_j and β_i and also to problem selection is Math Garden's system for scoring responses. Responses are scored according to Eq. (1):

$$S_{ij} = \frac{(2x_{ij} - 1)(d - rt_{ij})}{d} \tag{1}$$

Here S_{ij} denotes the score given to the response by user j to problem i, x_{ij} denotes the accuracy of the response (1 if correct, 0 if incorrect), d denotes the time allowed for a response, and rt_{ij} denotes the actual response time. Thus, S_{ij} is 1 (−1) for a correct (incorrect) response given immediately, 0.8 (−0.8) for a correct (incorrect) response given after 20% of the allowed time has elapsed, and so on. With this scoring system, the expected score for user j on problem i is given by Eq. (2) (Maris & van der Maas, 2012):

$$E(S_{ij}) = \frac{e^{2(\theta_j - \beta_i)} + 1}{e^{2(\theta_j - \beta_i)} - 1} - \frac{1}{\theta_j - \beta_i} \tag{2}$$

Again, θ_j and β_i denote the present ratings of user ability and problem difficulty, respectively.

New users' ability ratings θ_j are initialized based on their age, while problem difficulty β_i ratings are initialized based on estimated difficulty relative to other problems. After a given user responds to a given problem, the values of θ_j and β_i are then updated using a method invented by Elo (1978), in which the expected result (Eq. (2)) is compared to the actual result (Eq. (1)). This method is formalized in Eqs. (3) and (4):

$$\hat{\theta}_j = \theta_j + K_j(S_{ij} - E(S_{ij})) \tag{3}$$

$$\hat{\beta}_i = \beta_i + K_i(E(S_{ij}) - S_{ij}) \tag{4}$$

Here $\hat{\theta}_j$ and $\hat{\beta}_i$ denote the updated ratings of user ability and problem difficulty, respectively. The terms K_i and K_j determine the rate at which adjustment takes place, and themselves depend on estimates of the uncertainty of the current ratings of user ability and problem difficulty. The calculation of these terms is detailed in Klinkenberg et al. (2011). The effect of Eqs. (3) and (4) is that scores higher than expected lead to increases in ratings of user ability and decreases in ratings of problem difficulty, while scores lower than

expected lead to the opposite changes. Note that the same set of problem difficulty ratings is shared among all users, and similarly, the same set of all user ability ratings is shared among all problems.

When it is time for a new problem to be presented to a user, current ratings of that user's ability and the difficulty of all problems are used to select a new problem for which the given user is estimated to have about a 75% probability of answering correctly.[5] The estimated probability of a correct answer depends on user ability and problem difficulty ratings via a logistic function, as described in Eq. (5) (Klinkenberg et al., 2011):

$$P(x_{ij}) = \frac{e^{\theta_j - \beta_i}}{1 + e^{\theta_j - \beta_i}} \tag{5}$$

To obtain a problem for which a given user has approximately a 75% chance of answering correctly, the new item is selected whose difficulty rating is nearest to β_t, as determined by Eqs. (6) and (7). The 20 problems most recently presented to the same user are excluded from this selection.

$$P \sim N(0.75, 0.1) \tag{6}$$

$$\beta_t = \theta_j + \ln\left(\frac{P}{1-P}\right) \tag{7}$$

Eq. (6) says that probability P is drawn from a normal distribution with mean 0.75 and standard deviation 0.1. Eq. (7) says that the target difficulty rating of the new problem minus the user ability rating is an inverse logistic function of P. Thus, according to Eq. (5), P is the probability of a correct response to the new problem.

Appendix B

The problem difficulty ratings obtained in the study would only be meaningful if they had converged to relatively stable values by the end of data collection. The general ability of the Math Garden algorithm to converge on stable values was shown by Klinkenberg et al. (2011). To check whether such stability had been achieved in the present study, a simple linear regression was performed to predict final difficulty ratings using difficulty ratings obtained, for each problem, at the beginning of the last 200 trials involving that problem. The regression was significant, $F(1, 306) = 7300$, $p < .001$, adjusted $R^2 = .960$. The coefficient of earlier difficulty rating was near unity (1.015), while the model intercept was near zero (−0.034), indicating that the earlier difficulty ratings were nearly identical to the final difficulty ratings. Thus, difficulty ratings were highly stable over the last 200 trials for each problem.

It is important to establish not only that difficulty ratings were stable at the end of the study, but also that they changed substantially earlier in the study. Otherwise, the final difficulty ratings might simply reflect their initial values. To assess this possibility, two additional analyses were conducted. First, problems' difficulty ratings after their first 200 trials were regressed against their initial difficulty ratings at the beginning of the study. This regression was significant, $F(1, 306) = 108$, $p < .001$, $R^2 = .259$, but the amount of variance explained (25.9%) was substantially less than in the previous model (96.0%). Also, the coefficient of initial difficulty rating was not near unity (0.312), nor was the intercept near zero (4.87), indicating that difficulty ratings changed substantially over the first 200 trials, unlike the last 200 trials. As a second test, the final difficulty ratings were regressed separately against initial ratings and ratings after the first 200 trials. Both regressions were significant, $F(1, 306) = 90.5$, $p < .001$, adjusted $R^2 = .226$ for the regression using initial ratings and $F(1, 306) = 539$, $p < .001$, adjusted $R^2 = .637$ for the regression using ratings after the first 200 trials. However, the amount of variance explained (22.6% for initial difficulty ratings, 63.7% for ratings after the first 200 trials) was substantially less than that explained by regression against ratings at the beginning of the last 200 trials (96.0%). Thus, final difficulty ratings were much better predicted by difficulty ratings observed late in the study than by initial or early difficulty ratings. In sum, the stability observed at the end of the study was not a result of ratings adjusting slowly or not at all.

[5] While selecting problems with a target of 50% accuracy would yield more information for purposes of calculating user ability and problem difficulty ratings, 75% was preferred as a target in order to avoid discouraging users with low success rates. Math Garden's incorporation of response time into its scoring system allows reliable estimates of player ability levels despite this relatively high target accuracy. Math Garden users may elect to solve easier problems, in which case the target accuracy is set at 90%, or harder problems, in which case it is set at 60%.

Appendix C. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.cognition.2016.01.004.

References

Alibali, M. W., Phillips, K. M. O., & Fischer, A. D. (2009). Learning new problem-solving strategies leads to changes in problem representation. *Cognitive Development, 24*(2), 89–101. http://dx.doi.org/10.1016/j.cogdev.2008.12.005.

Alter, A. L., Oppenheimer, D. M., Epley, N., & Eyre, R. N. (2007). Overcoming intuition: Metacognitive difficulty activates analytic reasoning. *Journal of Experimental Psychology: General, 136*(4), 569–576. http://dx.doi.org/10.1037/0096-3445.136.4.569.

Anderson, J. R. (2005). Human symbol manipulation within an integrated cognitive architecture. *Cognitive Science, 29*(3), 313–341.

Anderson, J. R. (2009). *How can the human mind occur in the physical universe?* USA: Oxford University Press.

Bassok, M., Chase, V. M., & Martin, S. A. (1998). Adding apples and oranges: Alignment of semantic and formal knowledge. *Cognitive Psychology, 35*(2), 99–134. http://dx.doi.org/10.1006/cogp.1998.0675.

Bassok, M., Wu, L., & Olseth, K. L. (1995). Judging a book by its cover: Interpretative effects of content on problem-solving transfer. *Memory & Cognition, 23*(3), 354–367.

Bell, A., Fischbein, E., & Greer, B. (1984). Choice of operation in verbal arithmetic problems: The effects of number size, problem structure and context. *Educational Studies in Mathematics, 15*, 129–147. Retrieved from <http://link.springer.com/article/10.1007/BF00305893>.

Bell, A., Swan, M., & Taylor, G. (1981). Choice of operation in verbal problems with decimal numbers. *Educational Studies in Mathematics, 12*, 399–420. Retrieved from <http://link.springer.com/article/10.1007/BF00308139>.

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. http://dx.doi.org/10.1146/annurev-psych-113011-143823.

Briars, D., & Siegler, R. S. (1984). A featural analysis of preschoolers' counting knowledge. *Developmental Psychology, 20*(4), 607–618. http://dx.doi.org/10.1037//0012-1649.20.4.607.

Brissiaud, R., & Sander, E. (2010). Arithmetic word problem solving: A situation strategy first framework. *Developmental Science, 13*(1), 92–107. http://dx.doi.org/10.1111/j.1467-7687.2009.00866.x.

Bulloch, M. J., & Opfer, J. E. (2009). What makes relational reasoning smart? Revisiting the perceptual-to-relational shift in the development of generalization. *Developmental Science, 12*(1), 114–122. http://dx.doi.org/10.1111/j.1467-7687.2008.00738.x.

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*(1), 55–81. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/0010028573900042>.

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/S0364021381800298>.

Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educational Psychologist, 47*(3), 177–188. http://dx.doi.org/10.1080/00461520.2012.695709.

Crooks, N. M., & Alibali, M. W. (2013). Noticing relevant problem features: Activating prior knowledge affects problem solving by guiding encoding. *Frontiers in Psychology, 4*(884), 1–10. http://dx.doi.org/10.3389/fpsyg.2013.00884.

De Lima, R. N., & Tall, D. O. (2008). Procedural embodiment and magic in linear equations. *Educational Studies in Mathematics, 67*(1), 3–18. Retrieved from <http://link.springer.com/article/10.1007/s10649-007-9086-0>.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General, 122*(3), 371–396. Retrieved from <http://psycnet.apa.org/psycinfo/1993-44067-001>.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2010). Fortune favors the Bold (and the Italicized): Effects of disfluency on educational outcomes. *Cognition, 118*, 114–118. http://dx.doi.org/10.1016/j.cognition.2010.09.012.

Elo, A. (1978). *The rating of chess players, past and present*. New York, USA: Arco Publishers. Retrieved from <http://scholar.google.com/scholar?q=author:elo+1978+chess&btnG=&hl=en&as_sdt=0,31#1>.

Fischbein, E., Deri, M., Nello, M., & Marino, M. (1985). The role of implicit models in solving verbal problems in multiplication and division. *Journal for Research in Mathematics Education, 16*, 3–17. Retrieved from <http://www.jstor.org/stable/748969>.

Fischer, M. H., & Shaki, S. (2014). Spatial associations in numerical cognition – From single digits to arithmetic. *Quarterly Journal of Experimental Psychology (2006), 67*(8), 1461–1483. http://dx.doi.org/10.1080/17470218.2014.927515.

Fisher, K. J., Borchert, K., & Bassok, M. (2011). Following the standard form: Effects of equation format on algebraic modeling. *Memory & Cognition, 39*(3), 502–515. http://dx.doi.org/10.3758/s13421-010-0031-6.

Friedrich, R., & Friederici, A. (2009). Mathematical logic in the human brain: Syntax. *PLoS ONE*, e5599. Retrieved from <http://dx.plos.org/10.1371/journal.pone.0005599.g003>.

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development, 59*, 47–59. Retrieved from <http://www.jstor.org/stable/1130388>.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195–235). Cambridge, MA: MIT Press.

Gentner, D., & Toupin, C. (1986). Systematicity and surface similarity in the development of analogy. *Cognitive Science, 10*(3), 277–300. http://dx.doi.org/10.1016/S0364-0213(86)80019-2.

Goldstone, R. L., Landy, D., & Brunel, L. C. (2011). Improving perception to make distant connections closer. *Frontiers in Psychology, 2*(December), 1–10. http://dx.doi.org/10.3389/fpsyg.2011.00385.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2008). A well grounded education: The role of perception in science and mathematics. *Symbols, Embodiment, and Meaning*, 327–355.

Goldstone, R. L., Landy, D. H., & Son, J. Y. (2010). The education of perception. *Topics in Cognitive Science, 2*(2), 265–284. http://dx.doi.org/10.1111/j.1756-8765.2009.01055.x.

Herscovics, N., & Linchevski, L. (1994). A cognitive gap between arithmetic and algebra. *Educational Studies in Mathematics, 27*, 59–78. Retrieved from <http://link.springer.com/article/10.1007/BF01284528>.

Jansen, B. R. J., de Lange, E., & van der Molen, M. J. (2013). Math practice and its influence on math skills and executive functions in adolescents with mild to borderline intellectual disability. *Research in Developmental Disabilities, 34*, 1815–1824. Retrieved from <http://www.sciencedirect.com/science/article/pii/S0891422213000863>.

Jansen, B. R. J., Hofman, A., Straatemeier, M., van Bers, B., Raijmakers, M. E. J., & van der Maas, H. L. J. (2014). The role of pattern recognition in children's exact enumeration of small numbers. *British Journal of Developmental Psychology, 32*(2), 178–194. Retrieved from <http://onlinelibrary.wiley.com/doi/10.1111/bjdp.12032/full>.

Jansen, B. R. J., & Louwerse, J. (2013). The influence of experiencing success in math on math anxiety, perceived math competence, and math performance. *Learning and Individual Differences, 24*, 190–197. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1041608012001951>.

Jansen, A. R., Marriott, K., & Yelland, G. W. (2003). Comprehension of algebraic expressions by experienced users of mathematics. *The Quarterly Journal of Experimental Psychology, 56A*(1), 3–30. http://dx.doi.org/10.1080/02724980244000134.

Jansen, A. R., Marriott, K., & Yelland, G. W. (2007). Parsing of algebraic expressions by experienced users of mathematics. *European Journal of Cognitive Psychology, 19*(2), 286–320.

Jiang, M. J., Cooper, J. L., & Alibali, M. W. (2014). Spatial factors influence arithmetic performance: The case of the minus sign. *Quarterly Journal of Experimental Psychology, 67*, 1626–1642. http://dx.doi.org/10.1080/17470218.2014.898669.

Kamii, C., & Dominick, A. (1998). The harmful effects of algorithms in grades 1–4. In L. J. Morrow & M. J. Kenney (Eds.), *The teaching and learning of algorithms in school mathematics, 1998 yearbook* (pp. 130–140). Reston, VA: National Council of Teachers of Mathematics.

Keil, F. C. (1989). *Concepts, kinds, and conceptual development*. Cambridge, MA: MIT Press. Retrieved from <https://scholar.google.com/scholar?q=author:frank+author:keil&hl=en&as_sdt=0,49&as_ylo=1989&as_yhi=1989#1>.

Keil, F. C., & Batterman, N. (1984). A characteristic-to-defining shift in the development of word meaning. *Journal of Verbal Learning and Verbal Behavior, 23*(2), 221–236. http://dx.doi.org/10.1016/S0022-5371(84)90148-8.

Keil, F. C., Smith, W. C., Simons, D. J., & Levin, D. T. (1998). Two dogmas of conceptual empiricism: Implications for hybrid models of the structure of knowledge. *Cognition, 65*(2–3), 103–135. http://dx.doi.org/10.1016/S0010-0277(97)00041-3.

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. In B. H. Ross (Ed.). *Psychology of learning and motivation* (Vol. 58, pp. 117–165). Elsevier. http://dx.doi.org/10.1016/B978-0-12-407237-4.00004-9.

Kellman, P. J., Massey, C. M., & Son, J. Y. (2010). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science, 2*(2), 285–305. http://dx.doi.org/10.1111/j.1756-8765.2009.01053.x.

Kirshner, D. (1989). The visual syntax of algebra. *Journal for Research in Mathematics Education, 20*(3), 274–287. Retrieved from <http://www.jstor.org/stable/749516>.

Kirshner, D., & Awtry, T. (2004). Visual salience of algebraic transformations. *Journal for Research in Mathematics Education, 35*(4), 224. http://dx.doi.org/10.2307/30034809.

Klinkenberg, S., Straatemeier, M., & van der Maas, H. L. J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. *Computers & Education, 57*(2), 1813–1824. http://dx.doi.org/10.1016/j.compedu.2011.02.003.

Koedinger, K. R., & Anderson, J. R. (1990). Abstract planning and perceptual chunks: Elements of expertise in geometry. *Cognitive Science, 14*(4), 511–550. Retrieved from <http://linkinghub.elsevier.com/retrieve/pii/036402139090008K>.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The Knowledge-Learning-Instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*(5), 757–798. http://dx.doi.org/10.1111/j.1551-6709.2012.01245.x.

Landy, D. H. (2007). *Formal notations as diagrams of abstract structure*. Doctoral Dissertation, Indiana University.

Landy, D. H., Allen, C., & Zednik, C. (2014). A perceptual account of symbolic reasoning. *Frontiers in Psychology, 5*(April), 1–10. http://dx.doi.org/10.3389/fpsyg.2014.00275.

Landy, D. H., Jones, M. N., & Goldstone, R. L. (2008). How the appearance of an operator affects its formal precedence. In *Proceedings of the 30th annual conference of the Cognitive Science Society* (pp. 2109–2114).

Landy, D. H., & Goldstone, R. L. (2007a). Formal notations are diagrams: Evidence from a production task. *Memory & Cognition, 35*(8), 2033–2040.

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

Landy, D. H., & Goldstone, R. L. (2010). Proximity and precedence in arithmetic. *Quarterly Journal of Experimental Psychology, 63*(10), 1953–1968. http://dx.doi.org/10.1080/17470211003787619.

Linchevski, L., & Livneh, D. (1999). Structure sense: The relationship between algebraic and numerical contexts. *Educational Studies in Mathematics, 40*, 173–196.

Maris, G., & van der Maas, H. L. J. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika, 77*(4), 615–633. Retrieved from <http://link.springer.com/article/10.1007/s11336-012-9288-y>.

Martin, S. A., & Bassok, M. (2005). Effects of semantic cues on mathematical modeling: Evidence from word-problem solving and equation construction tasks. *Memory & Cognition, 33*(3), 471–478.

Maruyama, M., Pallier, C., Jobert, A., Sigman, M., & Dehaene, S. (2012). The cortical representation of simple mathematical expressions. *NeuroImage, 61*(4), 1444–1460. http://dx.doi.org/10.1016/j.neuroimage.2012.04.020.

McCrink, K., & Opfer, J. E. (2014). Development of spatial-numerical associations. *Current Directions in Psychological Science, 23*(6), 439–445. http://dx.doi.org/10.1177/0963721414549751.

McNeil, N. M. (2007). U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Developmental Psychology*. Retrieved from <http://psycnet.apa.org/journals/dev/43/3/687/>.

McNeil, N. M., & Alibali, M. W. (2004). You'll see what you mean: Students encode equations based on their knowledge of arithmetic. *Cognitive Science, 28*(3), 451–466. http://dx.doi.org/10.1016/j.cogsci.2003.11.002.

McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development, 76*(4), 883–899. http://dx.doi.org/10.1111/j.1467-8624.2005.00884.x.

Miller, K., Perlmutter, M., & Keating, D. (1984). Cognitive arithmetic: Comparison of operations. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(1), 46–60. Retrieved from <http://psycnet.apa.org/journals/xlm/10/1/46/>.

Monti, M., Parsons, L., & Osherson, D. (2012). Thought beyond language neural dissociation of algebra and natural language. *Psychological Science*. Retrieved from <http://pss.sagepub.com/content/23/8/914.short>.

Moore, A. M., & Ashcraft, M. H. (2015). Children's mathematical performance: Five cognitive tasks across five grades. *Journal of Experimental Child Psychology, 135*, 1–24. http://dx.doi.org/10.1016/j.jecp.2015.02.003.

Moyer, R. S., & Landauer, T. K. (1967). Time required for judgments of numerical inequality. *Nature, 215*, 1519–1520.

Novick, L. R. (1988). Analogical transfer, problem similarity, and expertise. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14*(3), 510–520. Retrieved from <http://www.ncbi.nlm.nih.gov/pubmed/2969945>.

Oppenheimer, D. M. (2008). The secret life of fluency. *Trends in Cognitive Sciences, 12*(6), 237–241. http://dx.doi.org/10.1016/j.tics.2008.02.014.

Ottmar, E., Landy, D. H., & Goldstone, R. L. (2012). Teaching the perceptual structure of algebraic expressions: Preliminary findings from the pushing symbols intervention. In *Proceedings of the 34th annual conference of the cognitive science society*. Retrieved from <http://cognitrn.psych.indiana.edu/rgoldsto/pdfs/pushingsymbols.pdf>.

Ottmar, E., Weitnauer, E., Landy, D., & Goldstone, R. L. (2015). Graspable mathematics: Using perceptual learning technology. In *Integrating touch-enabled and mobile devices into contemporary mathematics education*.

Piaget, J. (1952). *The origins of intelligence in children*. New York, NY: WW Norton & Co.

Rattermann, M. J., & Gentner, D. (1998). More evidence for a relational shift in the development of analogy: Children's performance on a causal-mapping task.

Cognitive Development, 13(4), 453–478MIT Press. Retrieved from <http://www.sciencedirect.com/science/article/pii/S088520149890003X>.

Richland, L. E., Stigler, J. W., & Holyoak, K. J. (2012). Teaching the conceptual structure of mathematics. *Educational Psychologist, 47*(3), 189–203.

Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology, 101*(3), 529–544. http://dx.doi.org/10.1037/a0014224.

Scheepers, C., Sturt, P., Martin, C. J., Myachykov, A., Teevan, K., & Viskupova, I. (2011). Structural priming across cognitive domains: From simple arithmetic to relative-clause attachment. *Psychological Science, 22*(10), 1319–1326. http://dx.doi.org/10.1177/0956797611416997.

Schneider, E., Maruyama, M., Dehaene, S., & Sigman, M. (2012). Eye gaze reveals a fast, parallel extraction of the syntax of arithmetic formulas. *Cognition, 125*(3), 475–490. http://dx.doi.org/10.1016/j.cognition.2012.06.015.

Sfard, A. (1991). On the dual nature of mathematical conceptions: Reflections on processes and objects as different sides of the same coin. *Educational Studies in Mathematics, 22*, 1–36.

Shrager, J., & Siegler, R. S. (1998). SCADS: A model of children's strategy choices and strategy discoveries. *Psychological Science, 9*(5), 405–410.

Siegler, R. S., & Stern, E. (1998). Conscious and unconscious strategy discoveries: A microgenetic analysis. *Journal of Experimental Psychology: General, 127*(4), 377–397.

Simons, D. J., & Keil, F. C. (1995). An abstract to concrete shift in the development of biological thought: The insides story. *Cognition, 56*(2), 129–163. http://dx.doi.org/10.1016/0010-0277(94)00660-D.

Star, J. R. (2005). Reconceptualizing procedural knowledge. *Journal for Research in Mathematics Education, 36*(5), 404–411.

Tall, D. O. (1995). Cognitive growth in elementary and advanced mathematical thinking. In *PME conference* (pp. 1–61). The Program committee of the 18th PME conference. Retrieved from <http://digilander.libero.it/leo723/materiali/algebra/dot1995b-pme-plenary.pdf>.

Tall, D. O. (2008). The transition to formal thinking in mathematics. *Mathematics Education Research Journal, 20*(2), 5–24. http://dx.doi.org/10.1007/BF03217474.

Van der Ven, S., van der Maas, H. L. J., Straatemeier, M., & Jansen, B. R. J. (2013). Visuospatial working memory and mathematical ability at different ages throughout primary school. *Learning and Individual Differences, 27*, 182–192. Retrieved from <http://www.sciencedirect.com/science/article/pii/S1041608013001192>.

Van Galen, M. S., & Reitsma, P. (2008). Developing access to number magnitude: A study of the SNARC effect in 7- to 9-year-olds. *Journal of Experimental Child Psychology, 101*(2), 99–113. http://dx.doi.org/10.1016/j.jecp.2008.05.001.

Varma, S., & Schwartz, D. L. (2011). The mental representation of integers: An abstract-to-concrete shift in the understanding of mathematical concepts. *Cognition, 121*, 363–385. http://dx.doi.org/10.1016/j.cognition.2011.08.005.

Vygotsky, L. S. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Wertheimer, M. (1938). Laws of organization in perceptual forms. Retrieved from <http://doi.apa.org/psycinfo/2007-10344-005>.

Wood, G., Willmes, K., Nuerk, H., & Fischer, M. (2008). On the cognitive link between space and number: A meta-analysis of the SNARC effect. *Psychology Science, 50*(4), 489–525. Retrieved from <http://doi.apa.org/?uid=2009-00781-003>.