An eye for relations: eye-tracking indicates long-term negative effects of operational thinking on understanding of math equivalence

Dana L. Chesney · Nicole M. McNeil ·
James R. Brockmole · Ken Kelley

Abstract Prior knowledge in the domain of mathematics can sometimes interfere with learning and performance in that domain. One of the best examples of this phenomenon is in students' difficulties solving equations with operations on both sides of the equal sign. Elementary school children in the U.S. typically acquire incorrect, operational schemata rather than correct, relational schemata for interpreting equations. Researchers have argued that these operational schemata are never unlearned and can continue to affect performance for years to come, even after relational schemata are learned. In the present study, we investigated whether and how operational schemata negatively affect undergraduates' performance on equations. We monitored the eye movements of 64 undergraduate students while they solved a set of equations that are typically used to assess children's adherence to operational schemata (e.g., $3 + 4 + 5 = 3 + \_\_$). Participants did not perform at ceiling on these equations, particularly when under time pressure. Converging evidence from performance and eye movements showed that operational schemata are sometimes activated instead of relational schemata. Eye movement patterns reflective of the activation of relational schemata were specifically lacking when participants solved equations by adding up all the numbers or adding the numbers before the equal sign, but not when they used other types of incorrect strategies. These findings demonstrate that the negative effects of acquiring operational schemata extend far beyond elementary school.

D. L. Chesney (✉) · N. M. McNeil · J. R. Brockmole
Department of Psychology, University of Notre Dame,
118 Haggar Hall,
Notre Dame, IN 46556, USA
e-mail: dlchesney@gmail.com

K. Kelley
Department of Management, University of Notre Dame, Notre
Dame, USA

Conventional wisdom suggests that the process of learning is a progressive march forward. As individuals gain domain knowledge, they become more proficient at operating within that domain. However, several decades of research indicate that this is not always the case. Prior domain knowledge sometimes interferes with the ability to operate successfully within that domain, particularly when task demands change (Bilalic, McLeod, & Gobet, 2008a, Bilalić et al. 2008b; Croskerry, 2003; Lippman, 1994; Lovett, & Anderson, 1996; Luchins, 1942; Wiley, 1998). One reason is that learners tend to rely on prior knowledge instead of encoding new information or generating new strategies. This *change resistance* has been shown in several areas of psychology (e.g., Allport, 1954; Diamond & Kirkham, 2005; Luchins, 1942; Munakata, 1998; Rescorla, 1996; Schauble, 1990; Wiley, 1998; Zelazo, Frye, & Rapus, 1996). In the present study, we focused on change resistance involving adherence to schemata that form through repeated experience in a given domain. We will refer to this type of change resistance as the *Einstellung effect* (Luchins, 1942), although it historically has been described using a variety of terms, including *habit* (James, 1890), *fixation* (Duncker, 1945; see also Maier, 1931), *mental set* (Wiley, 1998), and *rigidity* (Schultz & Searleman, 2002), among others. Our goal was to use eye-tracking methods to detect evidence of long-term, pernicious Einstellung effects in the domain of mathematics.

Einstellung effects

The Einstellung effect was classically demonstrated by Luchins's (1942) water jar problems. In these problems, individuals used water jars of known volumes (e.g., 18,

43, and 10 quarts) to construct a third volume (e.g., 5 quarts). After solving several problems that required a particular multistep solution strategy (e.g., $43 - 10 - 10 - 18 = 5$), problem solvers continued to employ this *set* strategy even when they later encountered problems for which that strategy was inefficient (e.g., finding 20 quarts given 23-, 49-, and 3-quart jars by $49 - 3 - 3 - 23$, rather than $23 - 3$) or even ineffective (e.g., finding 25 quarts given 28-, 76-, and 3-quart jars; see also Lovett & Anderson, 1996).

Einstellung effects have been linked to what is dubbed the *insight problem*: Individuals who are demonstrably capable of solving a problem can initially fail to see the solution—sometimes perceiving the problem as unsolvable—and will perseverate on ineffective solution strategies before an "Aha" moment of insight is reached and the correct solution strategy is discovered (Duncker, 1945; Ellis, Glaholt, & Reingold, 2011; Knoblich, Ohlsson, Haider, & Rhenius, 1999; Knoblich, Ohlsson, & Raney, 2001). Knoblich and colleagues (Knoblich et al., 1999; Knoblich et al., 2001) have suggested that difficulties achieving insight occur when initial conceptualizations of a problem—the mental representations dictated by prior knowledge—are incompatible with the correct solution. Consequently, these initial conceptualizations must be changed before correct solutions can be discovered. For example, if one conceptualizes water jar problems as problems that should be solved via the strategy $B - 2C - A$, this conceptualization must change before one can construct 25 quarts given 28-, 76-, and 3-quart jars.

Einstellung effects have been demonstrated not only in the short term, as set effects established under laboratory conditions (Lippman, 1994; Lovett, & Anderson, 1996; Luchins, 1942), but also in the long term, when domain experts' prior knowledge interferes with their ability to find optimal solutions (Bilalic et al., 2008a, 2008b; Croskerry, 2003; Wiley, 1998). This occurs in medicine. Doctors' initial framings of diagnostic problems can impair their ability to correctly diagnose conditions that fall outside these framings (Croskerry, 2003). It also occurs in chess. Chess masters fail to find the most efficient winning strategy when a less efficient but well-known winning strategy is available (Bilalic et al., 2008a, 2008b). Such pernicious Einstellung effects may not be limited to domain experts, such as physicians and chess masters, but, rather, may cause problems that affect everyday learning and cognition. Indeed, research suggests that long-term Einstellung effects can be observed in a domain that is ubiquitously taught in schools and is used in daily life: mathematics (e.g., Clement, Lochhead, & Monk, 1981; McNeil & Alibali, 2005b; McNeil, Rittle-Johnson, Hattikudur, & Petersen, 2010).

Evidence of pernicious Einstellung effects in mathematics

Even educated adults sometimes use previously learned, but currently inappropriate, strategies when solving math problems (Clement et al., 1981). For example, when translating word problems into equations, people often construct equations that mirror the syntax of the statements. Although this is effective if the statement is simple (e.g., "Billy has five apples and gets three more" correctly yields "$5 + 3 =$ __"), it often fails for more complex statements (e.g., "There are six times as many students as professors" incorrectly yields "$6 * S = P$," rather than "$S / 6 = P$" or "$S = 6 * P$"). Such mistakes can be explained as long-term Einstellung effects: Problem-solving ability is negatively affected when individuals think of problems in terms of previously learned but currently inappropriate schemata. These schemata interlink perceptual patterns with associated conceptualizations and relevant solution strategies (e.g., perceiving the number-embedded word problems above as problems involving direct syntax translation, which leads to one-to-one terminology matching in equation construction and, ultimately, the incorrect "$6 * S = P$" response).

McNeil and colleagues (McNeil & Alibali, 2005b; McNeil et al., 2010) have proposed that particular risk for the use of inappropriate strategies on some types of mathematics problems arises from the acquisition and entrenchment of incorrect or incomplete schemata involving the equal sign. Specifically, children may view the equal sign in terms of arithmetic operations, as a command to solve a problem (e.g., "add up all the numbers"), linked to an "$a + b = c$" format, with the "problem" on the left and the "answer" on the right (McNeil & Alibali, 2005a, 2005b). Research indicates that U.S. educational practices lead many children to initially acquire these *operational schemata* rather than correct *relational schemata* that treat the equal sign as a relational symbol conveying that two sides of an equation have the same value (Baroody & Ginsburg, 1983; Jacobs, Franke, Carpenter, Levi, & Battey, 2007; McNeil & Alibali, 2005a, 2005b).

In accordance with the growing body of research indicating that old concepts are never unlearned (see Ohlsson, 2009), the *change resistance account* of children's difficulties with algebra (McNeil & Alibali, 2005b; McNeil et al., 2010) suggests that acquiring operational schemata not only hinders later acquisition of relational schemata, but also can negatively affect mathematical performance even after relational schemata have been acquired. Such negative effects on performance would be predicted if the entrenchment of operational schemata yields Einstellung effects. Representing mathematical problems in terms of initially acquired operational schemata, rather than in terms of more recently acquired relational schemata, would be detrimental to performance when successful solving requires that problems be thought of relationally. For example, solving *mathematical equivalence*

problems—equations in which operations appear on both sides of the equal sign (e.g., 4 + 7 + 8 = 4 + _)—requires relational thinking. One must encode the problem in its entirety, conceptualize the equal sign as representing the equivalence of the two sides of the equation, and devise a strategy for equalizing the two sides. Such problems cannot be solved correctly if one relies on the operational schemata acquired in childhood. Such Einstellung effects would be particularly pernicious to higher mathematics, such as algebra, where relational thinking is required (Jacobs et al., 2007).

Several studies offer evidence that Einstellung effects indeed result from acquisition of operational schemata. Most 7- to 11-year-old children in the U.S. fail to solve math equivalence problems correctly (McNeil & Alibali, 2005b). Indeed, most use "typical arithmetic" solution strategies that reflect use of operational rather than relational schemata: They add all of the numbers (Add All: 23) or add all the numbers before the equal sign (Add to Equal Sign: 19) (McNeil & Alibali, 2005b). In contrast, children from countries whose early mathematics education does not promote operational schemata acquisition do not typically produce these errors (Li, Ding, Capraro, & Capraro, 2008). Furthermore, some studies have shown that younger children in the U.S., who likely have not yet solidified operational schemata, are more successful at solving these problems than are their older peers (McNeil, 2007).

Research suggests that the negative effects of acquiring operational schemata persist into adulthood, even in individuals who have acquired correct, relational schemata. Under time pressure, university students who completed elementary school in the U.S. use the same Add All and Add to Equal Sign strategies used by children, particularly after completing tasks intended to activate their operational schemata (McNeil & Alibali, 2005b; McNeil et al., 2010). One might speculate that the use of such typical arithmetic strategies is merely due to participants' overfamiliarity with calculating totals in everyday life (e.g., tallying a grocery bill) and subsequent overapplication of that highly practiced strategy under time pressure. However, evidence suggests that this is not the whole story. Although regular calculating of totals in everyday life may strengthen previously established operational schemata, it is not sufficient to cause use of typical arithmetic strategies on math equivalence problems. University students who completed elementary school in Asian countries, where educational practices do not typically result in operational schema acquisition, do not use typical arithmetic strategies on math equivalence problems under time pressure, even after completing tasks intended to activate operational schemata (McNeil et al., 2010). This suggests that the error of applying typical arithmetic strategies under time pressure results from the activation of previously acquired operational schemata, rather than from simply overapplying totaling strategies used in everyday life.

Motivation of the present study

The studies discussed above provide evidence that early acquisition of operational schemata yields long-term detrimental effects on mathematical performance. However, they are less useful for explaining why initial acquisition of operational schemata would interfere with future use of relational schemata to yield such effects. The studies discussed above share a common limiting factor in their design: They were able to detect the activation of operational or relational schemata during problem solving solely via the strategies used to solve the problems. As a result, they were only able to detect the activation of one schema at a time. Thus, these studies were unable to determine whether this impaired performance was indeed an Einstellung effect wherein possessing operational schemata was detrimental because they were sometimes activated *instead of*—rather than in addition to—relational schemata. An alternative explanation is that operational and relational schemata both became active when the adults were presented with math equivalence problems but that individuals sometimes proceeded to use operational typical arithmetic strategies rather than relational correct strategies. This would be a failure to correctly choose from among available strategies, rather than a typical Einstellung effect wherein activation of previously acquired schemata (operational schemata) inhibits activation of other, possibly more appropriate schemata (relational schemata). Indeed, there is evidence of dual schemata activation in children who are in the process of developing a correct understanding of math equivalence, resulting in exactly this kind of error in strategy choice. When asked to solve math equivalence problems, such children may use a typical arithmetic strategy while gesturing in a manner indicative of relational schemata activation (e.g., moving their left hand under the left side of an equation, pausing, and then moving their right hand under the right side of the equation; Goldin-Meadow, Alibali, & Church, 1993).

In order to determine whether such coactivation of relational and operational schemata also occurs in adults, one must be able to detect whether relational schemata are active during problem solving even when individuals ultimately use typical arithmetic strategies or, alternatively, whether typical arithmetic strategies are used when relational schemata activation is conspicuously absent during the solving process. Eye-movements are a potential source of such information. Eye-tracking, like gesture, can be used as a window to cognitive processes that may not be consciously accessible or apparent in task performance (Stephen, Boncoddo, Magnuson, & Dixon, 2009). Eye-tracking has been successfully used to detect mental activity both in studies of the Einstellung effect and in investigations of the solving of insight problems. When experiencing the Einstellung effect, expert chess players continued to look at features relevant to familiar, nonoptimal solutions at the expense of looking at

features relevant to optimal solutions (Bilalić et al., 2008b). Thus, mental activity characteristic of the Einstellung effect—fixation on well-known strategies to the detriment of other, currently more appropriate strategies—was detectable in the chess players' looking patterns. Further demonstrating that eye-tracking can detect and differentiate mental activity, studies on the *insight problem* found that participants showed increased fixation on relevant items preceding solving success (Ellis et al., 2011; see also Knoblich et al., 2001). Thus, eye-tracking was able to detect shifts in problem conceptualization needed to achieve insight *before* that insight manifested in a successful solution.

The utility of eye-tracking data is not limited to gaze fixations. Eye movements can also provide valuable information about cognitive processes related to problem solving, as demonstrated by Thomas and Lleras (2007). Participants were given Duncker's (1945) classic radiation problem. They viewed a diagram of four lasers arranged around a person with a tumor and were asked to determine how to destroy the tumor without harming the healthy tissue around it. The correct solution was to direct multiple low-powered lasers at the tumor from different trajectories, such that they passed through the healthy tissue without harm but intersected with an additive effect at the tumor to destroy it. Making eye motions that embodied this correct solution, such as looking back and forth between items that must interact (e.g., the tumor and multiple lasers), led to successful solving (Thomas & Lleras, 2007).

Given Thomas and Lleras's (2007) finding that patterns of eye movements can embody thought processes during problem solving, we used patterns of eye movements that embody relational thought processes as markers of relational schemata activation. Evidence from a successful gesture manipulation study informed our hypothesis about the types of eye movements most likely to be indicative of relational schemata activation. Cook, Mitchell, and Goldin-Meadow (2008) compared the learning of children who mimicked a relational gesture (moving their left hand under the left side of an equation, pausing, and then moving their right hand under the right side of the equation) with that of children who verbally repeated a relational phrase ("I want to make one side equal to the other side"). Note that in watching and performing the gesture, children in the gesture group moved their attention back and forth across the equal sign multiple times. These children gained more from instruction on how to solve math equivalence problems than did the children who mimicked the relational message in speech. Indeed, the movement of the gesturing children's attention back and forth across the equal sign may have played a significant role in the ultimate effectiveness of this intervention. Thus, we concluded that a looking pattern in which a participant's gaze traversed the problem—looking between the two sides of the equation—should be indicative of relational schemata activation; this gaze pattern physically embodies that relation.

In our study, we drew on Thomas and Lleras's (2007) method of using patterns of eye motion to detect mental activity. We monitored undergraduates' eye movements as they completed a timed math equivalence problem-solving task (see McNeil & Alibali, 2005b; McNeil et al., 2010). We were specifically interested in how often participants made eye movements indicative of relational schemata activation: looking back and forth between values on each side of the equal sign as if relating the two sides of the equation. We had no a priori assumptions about what would constitute a distinct operational looking pattern and, therefore, could not investigate operational looking patterns in this study.

Although such back-and-forth eye movements may reflect activation of relational schemata, they are neither necessary nor sufficient for use of a correct strategy. Considering the numbers in "$4 + 7 + 8 = 4 + \_\_$" in sequence, one can just as easily employ the correct strategy $4 + 7 + 8 − 4$ as the typical arithmetic strategy $4 + 7 + 8 + 4$. Furthermore, fixating on numbers out of sequence would not, in and of itself, prevent the use of a typical arithmetic strategy. One could add the numbers out of sequence (e.g., to make use of more easily recalled addend pairs: $4 + 4 = 8$, $8 + 8 = 16$, $16 + 7 = 23$). Indeed, no eye movements are necessary at all, given that all numbers can be read while fixating on a central point. Any strategy can be performed as long as the meanings and locations of all the numbers and symbols are known. Similar to gesture, the eye motion patterns observed are not *directly* tied to the use of a specific strategy. Rather, eye motions and strategy use are separate—although imperfect—indicators of thought processes. Just as relational gestures have been shown to indicate relational schemata activation even when the strategy used is not correct, it is also possible to consider the presence of back-and-forth looking patterns, where gaze traverses the equal sign, to indicate relational schemata activation even when the strategy used is not correct. We therefore predicted that participants who moved their eyes back and forth across the equal sign more often would be more likely to solve problems successfully, demonstrating more correct strategies and fewer typical arithmetic strategies than participants who made fewer traversing eye movements.

Method

Participants

Sixty-four undergraduates participated for course credit. Undergraduates attended a selective, mid-sized, private university in the Midwestern United States. Typical scores (25–75 percentile range) on the mathematics SAT for matriculating students at this university are between 680 and 770 (Grove, 2012).

Tasks

Since our goal was to determine whether undergraduates' use of typical arithmetic strategies to solve math equivalence problems is due to long-term Einstellung effects, our task of interest was the equation-solving task used in prior studies (described below). Participants also completed two additional tasks commonly used to assess understanding of math equivalence: equation encoding and defining the equal sign (described below; McNeil & Alibali, 2004, 2005a, 2005b). All possible task orders were counterbalanced. Thus, participants completed zero ($n = 20$), one ($n = 22$), or two ($n = 22$) tasks that had the potential to activate relational schemata prior to completing the equation-solving task. Participants who perform the solving task later in the sequence (*solving order*: first, second, or third) have more opportunities for their relational schemata to be activated. Thus, if use of typical arithmetic strategies to solve math equivalence problems is indeed due to Einstellung effects, participants who perform the solving task later in the sequence should be less likely to use correct strategies and more likely to use typical arithmetic strategies. No practice trials were given, since we were most interested in performance arising from participants' initial representations.

Equation-solving task

Participants solved eight equations, all in the format $a + b + c = d + \_\_$. Equations were adapted from previous studies examining cognitive processes involved in undergraduates' understanding of math equivalence (McNeil & Alibali, 2005b; McNeil et al., 2010). Our goal was to track participants' eye movements during the task, so we deviated from previous studies and asked participants to state the value of the blank aloud, rather than to write the value on an answer sheet. At the start of each trial, "READY," "SET," and "GO" were presented in the center of the screen sequentially for 500 ms each. Immediately following this sequence, the equation was presented in the center of the screen. Equations subtended 14° of visual angle horizontally and 0.7° vertically. The center-to-center distance between each character was 1.75°.

At the start of this task, participants saw the following instructions:

In this part of the experiment, you will solve some math problems. Solve the problems as quickly as possible, while still maintaining accuracy. Each problem will be presented for a brief period of time. After each problem disappears, say the answer aloud. After you have given your response, press a button on the game pad to see the next problem.

Consistent with previous studies (McNeil & Alibali, 2005b; McNeil et al., 2010), the first four of these trials were *timed*: The equation remained on the screen for 2 s before it disappeared, leaving a blank gray screen. The next four of these trials were *untimed*: The equation remained on the screen until participants (1) provided an answer and (2) advanced to the next trial with the press of a button on a handheld videogame controller. Untimed problems were preceded by identical instructions, except that "appears" replaced "disappears." All participants saw the same eight equations, in the same order, as indicated in the Appendix. If use of typical arithmetic strategies to solve math equivalence problems is indeed due to Einstellung effects, participants should be more likely to use correct strategies and less likely to use typical arithmetic strategies for problems later in the sequence (*problem number*: 1, 2, 3, or 4), since there would be more opportunity for their relational schemata to be activated. Furthermore, if looking back and forth across the equal sign indicates the activation of relational schemata, the number of traversing eye movements (*traversing score*) should positively predict use of correct strategies and negatively predict use of typical arithmetic strategies.

Encoding and defining tasks

The stimuli in the encoding and defining tasks were presented with the same apparatus and in the same format as stimuli in the equation-solving task. The encoding task included four math equivalence problems (see the Appendix) that were presented sequentially for 5 s each. After each, participants were asked to say aloud what they had seen. The defining task included a between-subjects manipulation of equal sign context. All participants viewed one of two problems with an arrow pointed toward the equal sign and were asked sequentially: "What is the name of this symbol?" "What does this symbol mean?" and "Can it mean anything else?" Half of the participants in each order condition ($n = 32$) viewed a math equivalence problem, and half ($n = 32$) viewed a traditional arithmetic problem. Although "=" symbolizes equality in both contexts, children viewing "=" embedded in a math equivalence problem have provided more relational definitions than have children viewing "=" embedded in a traditional problem (McNeil & Alibali, 2005a).

Apparatus

The equations were presented on a 21-in. CRT monitor with a screen refresh rate of 120 Hz. Throughout each trial, the spatial position of each observer's right eye was sampled at a rate of 1000 Hz by a tower-mounted EyeLink 2 K eye-tracking system (SR research, Inc.) running in pupil and corneal reflection mode, resulting in an average spatial accuracy of 0.15°. An eye movement was classified as a saccade if its amplitude exceeded 0.2° and either (1) its velocity exceeded 30 deg/s or (2) its acceleration exceeded 9,500 deg/s^2. All intersaccade instances were considered fixations. Chin and forehead rests stabilized head position

and kept viewing distance constant. Participants sat 81 cm from the screen, so that it subtended 28° of visual angle horizontally and 21° vertically. Stimuli consisted of equations (see the Appendix) and English language sentences (see the Encoding and Defining Task section above) presented in black text on a gray background. Verbal responses were recorded by a SONY® MP3 IC recorder.

Analyses and results

Overall performance

Encoding and defining tasks

Performance on the encoding task was near ceiling, with participants correctly reciting the math equivalence problems on 92.6 % of the trials. Performance on the defining task was not as good. Collapsing across context conditions, 33 (52 %) participants provided a relational definition (e.g., "equivalent to"). Of the remaining participants, 13 (20 %) provided operational definitions (e.g., "the sum of the numbers"), and 18 (28 %) provided definitions too unspecified to be deemed relational or operational (e.g., "equals"). (Four participants who provided relational definitions also offered operational definitions.) Participants viewing "=" in a traditional context were more likely to provide operational definitions (44 % vs. 9 %), $\chi^2(1, N = 64) = 9.69$, $p = .002$, than were participants viewing "=" in a math equivalence context. There was not a statistically significant difference for relational definitions (44 % vs. 59 %), $\chi^2(1, N = 64) = 1.56$, $p = .21$.

Solving tasks

Participants produced correct answers to 46.9 % of the equations on timed trials and 91.8 % on untimed trials. Of the 64 participants, 55 (86 %) made at least one error on the timed trials, and 14 (22 %) made at least one error on the untimed trials. Participants' strategies on timed trials were coded on the basis of their responses (McNeil & Alibali, 2005b). For example, for the problem $4 + 7 + 8 = 4 + \underline{\ \ }$, 15 is coded as Correct, indicating use of a relational, *correct* strategy. In contrast, 23 is coded as Add All, and 19 as Add to Equal Sign. We were particularly interested in the use of these operational, *typical arithmetic* strategies, since they are the incorrect strategies most often used by children. Participants used three other *identifiable incorrect* strategies: (1) Add Two, summing the wrong two numbers (e.g., 11), (2) Carry, stating one of the numbers on the left side of the equal sign (e.g., 7), and (3) Repeat, stating the number on the right side of the equal sign (e.g., 4). All other responses were coded as Other Incorrect. Responses were coded as reflecting a particular strategy involving calculation if they were within ±1 of the answer that would be achieved with that strategy. To establish interrater reliability, a second coder evaluated responses from 20 % of the participants. Agreement on response categorization was 100 %. Table 1 presents the distribution of strategies. Note that typical arithmetic strategies accounted for 62 % of undergraduates' incorrect strategies. Thus, undergraduates did not merely make calculation errors but, rather, solved equations using the incorrect strategies typically used by children.

As was predicted, participants performed better when the equation-solving task was performed later in the task sequence. Participants who completed the equation-solving task

Table 1 Distribution of strategies used to solve math equivalence problems on timed and untimed trials and the mean number of traversing eye movements made on such trials

Strategy	No. instances	Percentage of trials	Mean traversals [95 % confidence intervals]
Timed trials			
Correct	125	49 %	0.94 [0.78, 1.10]
Add All *(typical arithmetic)*	74	29 %	0.31 [0.16, 0.46]
Add to Equal *(typical arithmetic)*	4	2 %	0.25 [−0.24, 0.74]
Add Two *(identifiable incorrect)*	13	5 %	1.31 [0.75, 1.87]
Carry *(identifiable incorrect)*	2	1 %	0.50 [−0.48, 1.48]
Repeat *(identifiable incorrect)*	5	2 %	0.60 [−0.18, 1.38]
Other incorrect	33	13 %	0.52 [0.21, 0.83]
Total	256	100 %	0.70 [0.58, 0.82]
Untimed trials			
Correct	240	94 %	2.40 [2.16, 2.64]
Add All *(typical arithmetic)*	13	5 %	1.54 [0.97, 2.11]
Add Two *(identifiable incorrect)*	2	1 %	3.50 [2.52, 4.48]
Repeat *(identifiable incorrect)*	1	<1 %	9.00 [N/A]
Total	256	100 %	2.36 [2.13, 2.59]

first typically used fewer correct strategies on the timed problems ($N = 20$, $M = 1.00$ [out of 4], $SD = 1.17$, $CI_{.95} = [0.45, 1.55]$) than did those who completed the task second ($N = 22$, $M = 1.82$, $SD = 1.47$, $CI_{.95} = [1.17, 2.47]$), who in turn did worse than those who completed the task third ($N = 22$, $M = 2.95$, $SD = 1.09$, $CI_{.95} = [2.47, 3.44]$), $F_{model}(2, 61) = 12.9$, $p < .001$, $\eta^2_p = .30$; $F_{linear\ term}(1, 62) = 25.8$, $\widehat{\beta} = 0.98$, $SE = .19$, $p < .001$, $R^2 = .29$. There was a corresponding decrease in the use of typical arithmetic strategies when the equation-solving task was performed later in the task sequence (first, $M = 2.25$ [out of 4], $SD = 1.65$, $CI_{.95} = [1.48, 3.02]$; second, $M = 1.23$, $SD = 1.63$, $CI_{.95} = [0.50, 1.95]$; third, $M = 0.27$, $SD = 0.63$, $CI_{.95} = [−0.01, 0.55]$), $F_{model}(2, 61) = 10.8$, $p < .001$, $\eta^2_p = .26$; $F_{linear\ term}(1, 62) = 21.9$, $\widehat{\beta} = −.99$, $SE = .21$, $p < .001$, $R^2 = .26$. This is consistent with research suggesting that simple exposure to math equivalence problems leads to improvements in performance on tasks assessing understanding of math equivalence (McNeil et al., 2012). Use of correct strategies on the untimed equations was near ceiling for participants, regardless of the task order (first, $M = 3.45$ out of 4, $SD = 1.23$, $CI_{.95} = [2.87, 4.03]$; second, $M = 3.82$, $SD = 0.50$, $CI_{.95} = [3.60, 4.04]$; third: $M = 3.95$ $SD = 0.21$, $CI_{.95} = [3.86, 4.05]$). Correspondingly, use of typical arithmetic strategies was near floor, regardless of task order (first, $M = 0.55$, $SD = 1.23$, $CI_{.95} = [−0.03, 1.13]$; second, $M = 0.09$, $SD = 0.43$, $CI_{.95} = [−0.10, 0.28]$; third, $M = 0$, $SD = 0$, $CI_{.95} = $ [undefined]).

Although the *number* of tasks performed before equation solving predicted solving performance, the *kinds* of tasks performed did not. In fact, there were no statistically significant correlations among defining, encoding, or timed solving performance (all $ps > .10$), nor were there any statistically significant effects of the kind(s) of task(s) that came before timed solving (all $ps > .10$). For example, although the context of the equal sign in the defining task affected participants' definitions, it did not significantly affect performance on timed solving trials among participants who completed the solving task after the defining task. This held for typical arithmetic strategy use (traditional arithmetic context, $M = .22$, $CI_{.95} = [0.05, 0.39]$; equivalence context, $M = .16$, $CI_{.95} = [0.00, 0.31]$), $t(30) = 0.53$, $p = .60$, correct strategy use (traditional, $M = .59$, $CI_{.95} = [0.41, 0.78]$; equivalence, $M = .63$, $CI_{.95} = [0.46, 0.79]$), $t(30) = −0.25$, $p = .80$, and traversing eye movements (traditional, $M = .72$, $CI_{.95} = [0.44, 1.00]$; equivalence, $M = .80$, $CI_{.95} = [0.53, 1.07]$), $t(30) = −0.39$, $p = .70$. Thus, these factors were not included in our models of timed solving performance.

Eye movement patterns predict strategy use

Given that there was a ceiling effect on untimed solving, we limited our analyses of the eye movement patterns to timed solving. For each equation, we tallied the number of times participants' gaze traversed the equation (i.e., when gaze fixation moved from items on one side of the equal sign to items on the other side). Recall that we consider looking between sides of the equation in this fashion to indicate that participants are *relating* the two sides of the problem. We included both cases where fixation moved from one side to the other directly or via the equal sign in this tally. Table 2 displays examples of participants' looking behaviors, along with their corresponding traversing scores for the problem $7 + 9 + 6 = 8 + $ __. Participants' average traversing score was 0.70 per equation ($N = 64$, $SD = .55$, $CI_{.95} = [0.57, 0.84]$).

Recall our hypotheses that traversing score would be positively associated with use of a correct strategy and negatively associated with use of a typical arithmetic strategy. Consistent with these predictions, the mean number of traversing eye movements was lower on trials where typical arithmetic strategies were used than on trials where correct strategies were used (see Table 1). We formally tested our hypotheses using a mixed-effects generalized linear model with a logit link function (i.e., a repeated measures logistic regression model) using the lme4 (Bates, Maechler, & Bolker, 2011) package in the R language and environment for statistical computing (R Development Core Team, 2012). In each model, the dichotomous outcome variable of correct strategy (yes/no) was modeled as the log odds (i.e., logit).

We used a model-building strategy. First, we constructed a model (model 1) in which only an intercept was used, both as a fixed effect and as a random effect (see Rabe-Hesketh & Skrondal, 2012). The fixed effect represents the average value across individuals, and the random effect is the

Table 2 Examples of participants' looking behaviors along with corresponding traversing scores for the equation $7 + 9 + 6 = 8 + $ __

Sample sequences of items fixated in the equation *7 + 9 + 6* = **8 +** _	Traversing score
First +, 9	0
6, 9, first +, 9, 6, **8**	1
6, 9, 6, **8**	1
6, **8**, =, **8**, *second +, 6, 9, second +, 9, 6*	2
Second +, 9, 6, **8**, *6,* **8**	3
9, second +, =, *6,* **third +, 8**, =, *6,* **8**, =, **8**, =	3

For clarity, items on the left side of the equal sign are presented here in *italics*, while items on the right side of the equal sign are in **bold**. The traversing score is calculated by tallying the number of times participants shifted their gaze from items on one side of the equal sign to items on the other side (traversals; e.g., "*6,* **8**" or "**8**, *9*"). The "=" is not considered to belong to a particular side. As such, it is *not* considered a traversal if a participant's gaze shifts back and forth between it and items on a given side and the equal sign (e.g., "**8**, =, **8**" is not a traversal). However, it is considered a traversal if a participant shifts his or her gaze from one side of the equation to the other *by way of* the equal sign (e.g., "**8**, =, *6*" is a traversal)

individual-specific deviation from the fixed effect. All subsequent models included this intercept. Next, we fitted a model (model 2) that included *solving order* (whether the equation solving task was performed first, second, or third), so as to assess how response patterns varied depending upon whether participants performed the solving task earlier or later in the sequence (orders 1st–3rd were coded as 0–2, so the intercept could be interpreted as the log odds [logit] that participants who performed the solving task first would use a correct strategy). Then we fitted a model (model 3) that included *problem number*, so as to assess how response patterns varied across time as the four timed equations were sequentially presented (Equations 1–4 were coded as 0–3, so the intercept could be interpreted as the log odds [logit] of the first problem being answered via a correct strategy). Then we fitted a model (model 4) that included *traversing score* as a time-varying covariate. Then we fitted a model (model 5) that included both solving order and problem number and a model (model 6) that included both solving order and traversing score. Then we fitted a model (model 7) that included solving order, problem number, and traversing score. Finally, we fitted a model (model 8) that included solving order, problem number, traversing score, and the possible interactions between and among them. Note that model 1 is a special case (i.e., nested) of all higher numbered models; models 2 and 3 are special cases of model 5; models 2 and 4 are special cases of model 6; models 5 and 6 are special cases of model 7; and model 7 is a special case of model 8. The results of this model-building approach are presented in Table 3.

The modeling results were consistent with our predictions. The "best" fitting model was model 7, which included main effects of traversing score, problem number, and solving order. This model provides a statistically significant improvement over models 1–6. Traversing score, problem number, and solving order all positively predicted use of a correct strategy: traversing score, $\widehat{\beta} = 0.76$, $z = 3.15$, $p = .002$; problem number, $\widehat{\beta} = 1.13$, $z = 5.84$, $p < .001$; solving order, $\widehat{\beta} = 1.99$, $z = 4.87$, $p < .001$. Including interactions (model 8) did not offer a statistically significant improvement over model 7. Thus, the interactions are not retained, and our final model (model 7) included only intercept, solving order, problem number, and traversals as predictors. Recall that in a logistic regression model, the regression coefficients represent the change in the log odds of success for a one-unit change in the corresponding regressor variable. We have transformed this log odds value into a probability for easier interpretation of the figures. As is shown in Fig. 1, the model-implied probability of using a correct equation-solving strategy increases as a function of traversing score, problem number, and solving order.

We arrived at similar conclusions when we used the same model-building approach to model use of a typical arithmetic strategy (yes/no). The "best" fitting model was again model

7. Traversing score, problem number, and solving order all negatively predicted use of a typical arithmetic strategy: traversing score, $\widehat{\beta} = -1.33$, $z = -3.15$, $p = .002$; problem number, $\widehat{\beta} = -0.61$, $z = -2.43$, $p = .015$; solving order, $\widehat{\beta} = -3.46$, $z = -3.98$, $p < .001$ (see Table 4). As is shown in Fig. 2, the model-implied probability of solving an equation with a typical arithmetic strategy decreases as a function of traversing score, problem number, and solving order.

Importantly, we tested to make sure traversing score was not simply a proxy for general attention to the equation during the brief presentation time. We modeled the continuous outcome variable *total time spent looking at the equation* using a mixed-effects general linear model. We found no evidence that traversing score had an effect on the total time spent looking at the equation in a model including solving order, problem number, and traversing score as predictors, $\widehat{\beta}_{traversals} = -32.99$, $z = -0.85$, $p = .39$. Indeed, the model that included traversing score was not significantly different from the simpler model that included just solving order and problem number as predictors, $\chi^2(1) = 0.72$, $p = .40$. We also performed analyses to determine whether focusing on the equal sign was positively correlated with correct strategy use on timed trials. It was not. Indeed, the opposite pattern was seen, $r = -.13$, $p = .03$. Additionally, longer fixation on the equal sign was related to use of typical arithmetic strategies, although this was not statistically significant, $r = .10$, $p = .12$. Possibly, participants whose operational schemata are activated find the equal sign's nonfinal location surprising. Regardless, it is apparent that simply looking at the equal sign was not sufficient to activate relational schemata.

Lack of traversing specifically predicts use of typical arithmetic strategies

We sought to determine the reason why traversing eye movements were positively correlated with correct strategy use and negatively correlated with typical arithmetic strategy use. It may be that relational schemata activation (as detected by traversing eye movements) is most prevalent with correct strategy use and is infrequent when any other strategies are used. However, a second possibility is that relational schemata activation is specifically lacking when typical arithmetic strategies (i.e., Add All, Add To Equal Sign) are used but is prevalent when correct and other incorrect strategies (i.e., Repeat, Carry, Add Two) are used. This second possibility is of particular interest, since it would demonstrate a key feature of the Einstellung effect: Individuals become entrenched in previously learned thought patterns at the expense of more beneficial ways of thinking. We tested these possibilities using the same model building approach described above, with two variations.

Table 3 Results and comparisons of models predicting the log odds (logit) of trials being solved with correct strategies from solving order, problem number and traversing eye movements on timed solving trials

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Factors: Model-derived parameters [95 % confidence intervals]								
Intercept	-0.11 [-0.65, 0.44]	-1.60*** [-2.38, -0.83]	-1.91*** [-2.84, -0.98]	-0.64* [-1.25, -0.03]	-4.00*** [-5.31, -2.70]	-2.17*** [-3.01, -1.32]	-4.47*** [-5.85, -3.09]	-4.92*** [-7.16, -2.68]
Solving order		1.46*** [0.87, 2.05]			2.03*** [1.20, 2.86]	1.48*** [0.90, 2.07]	1.99*** [1.19, 2.79]	2.45*** [1.00, 3.89]
Problem number			1.16*** [0.80, 1.52]		1.17*** [0.80, 1.54]		1.13*** [0.75, 1.51]	1.27** [0.36, 2.16]
Traversals				0.75*** [0.35, 1.15]		0.75*** [0.36, 1.14]	0.76** [0.28, 1.23]	0.94 [-.64, 2.53]
Solving order × problem number								-0.19 [-0.82, .44]
Solving order × traversals								-0.30 [-1.42, 0.81]
Problem number × traversals								-0.04 [-0.73, 0.81]
Solving order × problem number × traversals								-0.07 [-0.56, 0.69]
SD of random effect	1.86	1.41	2.70	1.82	2.13	1.34	1.96	1.92
AIC	252.4	232.6	211.1	241.0	191.2	220.1	183	189.7
BIC	259.5	243.2	221.7	251.7	205.4	234.2	200.7	221.6
Log likelihood	-124.2	-113.3	-102.5	-117.5	-91.6	-106.0	-86.5	-85.86
Deviance	248.4	226.6	205.1	235.0	183.2	212.1	173	171.7
Model Comparisons								
Compared to 1		$\chi^2(1) = 21.8$***	$\chi^2(1) = 43.3$***	$\chi^2(1) = 13.4$***	$\chi^2(2) = 65.2$***	$\chi^2(2) = 36.3$***	$\chi^2(3) = 75.4$***	$\chi^2(7) = 76.7$***
Compared to 2					$\chi^2(1) = 43.4$***	$\chi^2(1) = 14.5$***	$\chi^2(2) = 53.6$***	$\chi^2(6) = 54.8$***
Compared to 3					$\chi^2(1) = 21.9$***		$\chi^2(2) = 32.1$***	$\chi^2(6) = 33.4$***
Compared to 4						$\chi^2(1) = 23.0$***	$\chi^2(2) = 62.0$***	$\chi^2(6) = 63.3$***
Compared to 5							$\chi^2(1) = 10.2$**	$\chi^2(5) = 11.5$*
Compared to 6							$\chi^2(1) = 39.1$***	$\chi^2(5) = 40.3$***
Compared to 7								$\chi^2(4) = 1.3$

$p < .05$ represents a rejection of the null hypothesis that a factor does not significantly predict responding (top: Parameters) or that the models fit equally well (bottom: Model Comparisons). Problem numbers were coded as 0–3 rather than 1–4, and solving orders were coded as 0–3 rather than 1–3 for modeling purposes. The estimation method used for the analyses was a Gaussian Hermite approximation to the log-likelihood with 100 points per axis. In addition to the models described here, we also ran all models including two main effects and their interaction, as well as a three main effect model including all two-way interactions but excluding the three-way interaction. All models including interactions failed to show improvement over their corresponding main-effects-only models. Including the three-way interaction also failed to affect the conclusions

* $p < .05$

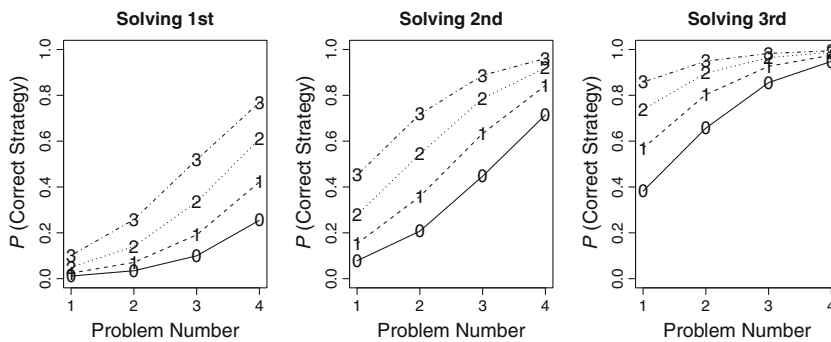** $p < .01$

*** $p < .001$

Fig. 1 The model-implied probability of an individual with a random effect (intercept) equal to the mean of the random effect distribution using a correct solving strategy as a function of solving task order (paneled), problem number (1st–4th, *x*-axis) and traversing score (separate lines, which are denoted with the number of traversals) for timed solving trials. Traversing score refers to the number of times participants' gazes moved from one side of the equation to the other (see Table 2)

In the first variation, we modeled how solving order, problem number, and traversing predicted use of a typical arithmetic strategy (yes/no), while excluding trials on which correct strategies were used. Results indicate that relational schemata were less activated when participants used typical arithmetic strategies than when they used other incorrect strategies: Traversing negatively predicted typical arithmetic strategy use, $\widehat{\beta}_{traversals} = -0.92$, $z = -2.45$, $p = .014$ (see Table 5). In the second variation, we modeled use of a correct strategy (yes/no), while excluding trials on which typical arithmetic strategies were used. We found no evidence to suggest that relational schemata were more activated when participants used correct strategies than when they used other, incorrect strategies: Traversing consistently failed to be a statistically significant predictor ($p > .07$; see Table 6)

Although these results are consistent with the idea that relational schemata activation is lacking only when incorrect, typical arithmetic strategies are used, one concern might be that cases coded as other incorrect strategies may simply be failed attempts at using correct strategies. Indeed, more than half these trials (33 of 53) were coded as "Other Incorrect." To ameliorate these concerns, we separately analyzed the 20 instances of participants using identifiable incorrect strategies that were not typical arithmetic strategies (i.e., Add Two, Carry, Repeat). These cases had a mean traversing score of 1.05 (CI$_{.95}$ = [0.61, 1.49]). This was not statistically different, $t(143) = 0.462$, $p = .644$ (Cohen's $d = 0.11$, CI$_{.95}$ = [−.36, .58]), from the traversing scores on trials on which correct strategies were used ($M = 0.94$, CI$_{.95}$ = [0.78, 1.10], $N = 125$), but they were significantly greater than, $t(96) = 4.122$, $p < .001$ (Cohen's $d = 1.03$, CI$_{.95}$ = [0.52, 1.54]) those on trials on which typical arithmetic strategies were used ($M = 0.31$, CI$_{.95}$ = [0.17, 0.45], $N = 78$). Confidence intervals for Cohen's d were calculated using the MBESS R package (Kelley, 2007a, 2007b). These results suggest that traversing eye movements were particularly lacking when participants used typical arithmetic strategies.

Discussion

Operational schemata yield Einstellung effects

Our results replicate and extend prior findings (McNeil & Alibali, 2005b; McNeil et al., 2010) suggesting that operational schemata negatively affect educated adults. Under time pressure, 47 % of the undergraduates participating in our study not only made errors, but also made errors consistent with the use of incorrect operational schemata. When faced with a math equivalence problem (e.g., 4 + 7 + 8 = 4 + __), participants often added all the numbers (e.g., 23) or added the numbers preceding the equal sign (e.g., 19), rather than correctly balancing the equation (e.g., 15). These are the same kinds of errors that U.S. children make before they acquire a relational understanding of the equal sign (McNeil & Alibali, 2005a, 2005b). Strikingly, our undergraduates used typical arithmetic strategies spontaneously, without any deliberate attempt on our part to activate their operational schemata. Indeed, participants who performed the solving task first and were thus unprimed were the most likely to use typical arithmetic strategies. Six participants even persisted in making these errors on untimed trials. Thus, our findings lend further support to the change resistance account (McNeil & Alibali, 2005b; McNeil et al., 2010): Acquisition of operational schemata yields Einstellung effects that have long-term negative effects on individuals' performance on tasks requiring one to think relationally.

Although our data supported our hypotheses, they do not, by themselves, rule out the previously discussed possibility that participants' incorrect use of typical arithmetic strategies results from overapplying the totaling strategies commonly used in everyday life (e.g., tallying a grocery bill). However, considering previous evidence that adults do not use typical arithmetic strategies if their elementary school experiences did not originally lead them to acquire operational schemata (McNeil et al., 2010), parsimony leads us to conclude that use of typical arithmetic strategies results from early acquisition and subsequent activation of operational schemata.

Table 4 Results and comparisons of models predicting the log odds (logit) of trials being solved with typical arithmetic strategies from solving order, problem number, and traversing eye movements on timed solving trials

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Factors: Model-derived parameters [95 % confidence intervals]								
Intercept	-2.34*** [-3.49, -1.19]	0.83 [-0.72, 2.37]	-1.64* [-3.06, -0.22]	-1.56* [-2.84, -0.28]	1.68* [0.05, 3.30]	1.78* [0.06, 3.50]	2.66** [0.70, 4.62]	3.06** [0.80, 5.32]
Solving order		-3.04*** [-4.52, -1.57]			-2.83*** [-4.20, -1.45]	-3.23*** [-4.80, -1.66]	-3.46*** [-5.16, -1.75]	-3.92*** [-5.96, -1.88]
Problem number			-0.71** [-1.15, -0.27]		-0.66* [-1.08, -0.24]		-0.61* [-1.10, -0.12]	-0.87 [-1.74, 0.00]
Traversals				-1.53*** [-2.36, -0.70]		-1.48*** [-2.28, -0.67]	-1.33*** [-2.15, -0.50]	-1.16 [-3.40, 1.08]
Solving order × problem number								0.30 [-0.52, 1.11]
Solving order × traversals								-0.29 [-2.56, 1.99]
Problem number × traversals								0.12 [-1.38, 1.14]
Solving order × problem number × traversals								0.18 [-1.06, 1.41]
SD of random effect	4.00	3.35	4.58	4.25	3.22	3.58	3.86	3.97
AIC	160.3	141.2	151.7	143.7	133.0	125.0	120.4	127.0
BIC	167.4	151.8	162.4	154.3	147.2	139.2	138.2	158.9
Log likelihood	-78.2	-67.6	-72.9	-68.8	-62.5	-58.5	-55.2	-54.5
Deviance	156.3	135.2	145.7	137.7	125.0	117.0	110.4	109.0
Model comparisons								
Compared to 1		$\chi^2(1) = 21.1$***	$\chi^2(1) = 10.6$**	$\chi^2(1) = 18.6$***	$\chi^2(2) = 31.3$***	$\chi^2(2) = 39.3$***	$\chi^2(3) = 45.9$***	$\chi^2(7) = 47.3$***
Compared to 2					$\chi^2(1) = 10.2$**	$\chi^2(1) = 18.2$***	$\chi^2(2) = 24.8$***	$\chi^2(6) = 26.2$***
Compared to 3					$\chi^2(1) = 20.7$***		$\chi^2(2) = 35.3$***	$\chi^2(6) = 36.7$***
Compared to 4						$\chi^2(1) = 20.7$***	$\chi^2(2) = 27.2$***	$\chi^2(6) = 28.7$***
Compared to 5							$\chi^2(1) = 14.5$**	$\chi^2(5) = 16.0$**
Compared to 6							$\chi^2(1) = 6.6$*	$\chi^2(5) = 8.0$
Compared to 7								$\chi^2(4) = 1.4$

$p < .05$ represents a rejection of the null hypothesis that a factor does not significantly predict responding (top: Parameters) or that the models fit equally well (bottom: Model Comparisons). Problem numbers were coded as 0–3 rather than 1–4, and solving orders were coded as 0–3 rather than 1–3 for modeling purposes. The estimation method used for the analyses was a Gaussian Hermite approximation to the log-likelihood with 100 points per axis. In addition to the models described here, we also ran all models including two main effects and their interaction, as well as a three main effect model including all two-way interactions but excluding the three-way interaction. All models including interactions failed to show improvement over their corresponding main-effects-only models. Including the three-way interaction also failed to affect the conclusions
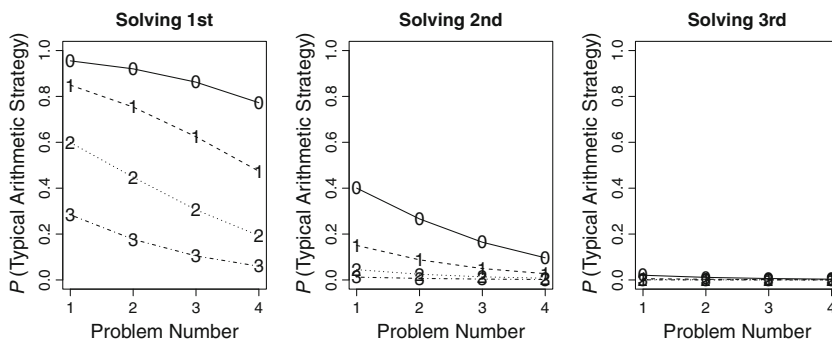
* $p < .05$

** $p < .01$

*** $p < .001$

Fig. 2 The model-implied probability of an individual with a random effect (intercept) equal to the mean of the random effect distribution solving an equation with a typical arithmetic strategy as a function of solving task order (paneled), problem number (1st–4th, *x*-axis), and traversing score (separate lines, which are denoted with the number of traversals) for timed solving trials. Traversing score refers to the number of times participants' gazes moved from one side of the equation to the other (see Table 2)

Our data further add to the literature by providing insights into how operational schemata yield long-term Einstellung effects. As was previously discussed, possessing operational schemata might cause decrements in mathematical performance if they are activated instead of (rather than in addition to) relational schemata. In such cases, individuals' initial problem representations will be solely operational, linked to operational problem formats, strategies, and concepts of the equal sign. Consequently, individuals will need to alter these initial representations—much as problem solvers need to do when faced with insight problems (Knoblich et al., 1999)—before they can correctly solve problems that require relational schemata. Analyses of our participants' eye movements suggest that this is the case.

Specifically, traversing eye movements were conspicuously absent when participants used typical arithmetic strategies. This finding could be explained if traversing eye movements specifically resulted from the use of correct strategies, rather than from the activation of relational schemata in general. However, that explanation was not supported by the data, since no differences in traversing score were detected between trials on which correct strategies were used and trials on which identifiable, incorrect, but *not* specifically typical arithmetic strategies (e.g., Add Two, Carry, Repeat) were used. Thus, traversing eye movements are not merely a marker of correct strategy use. This finding indicates that participants tended to use typical arithmetic strategies when relational schemata were *not* activated and, thus, participants did not link their initial representations of the problems to relational concepts and correct strategies.

We conclude that typical arithmetic strategy use occurs when operational schemata initially become activated instead of relational schemata. Consequently, initial representations of the problems are solely operational and must change before correct strategies can be accessed. This need to change initial representations is also consistent with our finding that participants who solved math equivalence problems later in the task sequence used more correct strategies and fewer typical

arithmetic strategies. The more math equivalence tasks participants completed before solving the math equivalence problems, the more chances they had to achieve the insight needed to change their initial representations to include relational schemata. Once participants achieve a moment of insight that activates their relational schemata, these schemata should remain active, facilitating correct strategy use when they perform later tasks (cf. Knoblich et al., 1999).

The possible perniciousness of operational Einstellung effects

Our participants demonstrated long-term Einstellung effects on their performance in mathematics, a domain of knowledge that is a mainstay of compulsory education in the U.S. Thus, our findings indicate that negative effects of domain knowledge, which previously have been observed in doctors (Croskerry, 2003) and chess masters (Bilalic et al., 2008a, 2008b), may affect a large proportion of the population. Such Einstellung effects might be particularly pernicious when there is no immediate feedback that one has used an incorrect strategy. Consider, for example, a water jar problem for which the set strategy is effective but inefficient. A solver can find 20 quarts given a 23-, 49-, and 3-quart jug via the strategy $49 - 3 - 3 - 23$, although the solution $23 - 3$ is more efficient. Similarly, when chess masters use familiar rather than efficient strategies, the problems still appear to be solved. Lacking immediate feedback on the efficacy of their strategies, they have little motive to change how they have represented the problem. Undergraduates' use of typical arithmetic strategies to solve math equivalence problems likewise lacks immediate feedback. When a solver uses the incorrect Add All strategy to solve $4 + 7 + 8 = 4 + \_\_$, he or she nevertheless arrives at a solution: 23. Unless informed by an outside source, the solver has no reason to believe that this solution is incorrect and, thus, has no motivation to alter the initial representation of the problems. Unfortunately, lack of immediate feedback is a common characteristic of the formal school

Table 5 Results and comparisons of models predicting the log odds (logit) of trials being solved with typical arithmetic strategies from solving order, problem number, and traversing eye movements on timed solving trials, excluding cases where a correct strategy was used

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Factors: Model-derived parameters [95 % confidence intervals]								
Intercept	−0.01 [−0.89, 0.88]	1.73** [0.46, 3.01]	−0.06 [−1.03, 0.92]	0.50 [−0.46, 1.45]	1.69* [0.33, 3.05]	2.21** [0.88, 3.55]	2.06** [0.68, 3.43]	2.44** [0.75, 4.13]
Solving order		−2.06*** [−3.22, −0.89]			−2.04*** [−3.20, −0.87]	−2.02*** [−3.17, −0.88]	−1.97*** [−3.08, −0.85]	−2.23** [−3.67, −0.79]
Problem number			−0.06 [−0.39, 0.51]		0.03 [−0.44, 0.50]		0.13 [−0.38, 0.63]	−0.17 [−1.03, 0.68]
Traversals				−0.92* [−1.66, −0.19]		−0.92* [−1.65, −0.18]	−0.93* [−1.67, −0.19]	−0.54 [−2.93, 1.85]
Solving order × problem number								0.22 [−0.55, 0.99]
Solving order × traversals								−1.10 [−3.82, 1.62]
Problem number × traversals								−0.16 [−1.56, 1.24]
Solving order × problem number × traversals								0.64 [−0.85, 2.14]
SD of random effect	2.66	2.43	2.61	2.60	2.40	2.36	2.29	2.46
AIC	102.3	92.2	104.2	98.0	94.2	88.0	89.8	95.0
BIC	108.0	100.9	112.9	106.7	105.7	99.5	104.1	120.8
Log likelihood	−49.1	−43.1	−49.1	−46.0	−43.1	−40.0	−39.9	−38.5
Deviance	98.3	86.2	98.2	92.0	86.2	80.0	79.8	77.0
Model comparisons								
Compared to 1		$\chi^2(1) = 12.0$***	$\chi^2(1) = 0.1$	$\chi^2(1) = 6.2$*	$\chi^2(2) = 12.1$**	$\chi^2(2) = 18.3$***	$\chi^2(3) = 18.5$***	$\chi^2(7) = 21.3$**
Compared to 2					$\chi^2(1) = 0.0$	$\chi^2(1) = 6.3$*	$\chi^2(2) = 6.5$*	$\chi^2(6) = 9.3$
Compared to 3					$\chi^2(1) = 12.0$***		$\chi^2(2) = 18.5$***	$\chi^2(6) = 21.3$**
Compared to 4						$\chi^2(1) = 12.1$***	$\chi^2(2) = 12.3$**	$\chi^2(6) = 15.1$**
Compared to 5							$\chi^2(1) = 6.5$*	$\chi^2(5) = 9.3$
Compared to 6							$\chi^2(1) = 0.2$	$\chi^2(5) = 3.0$
Compared to 7								$\chi^2(4) = 2.8$

$p < .05$ represents a rejection of the null hypothesis that a factor does not significantly predict responding (top: Parameters) or that the models fit equally well (bottom: Model Comparisons). Problem numbers were coded as 0–3 rather than 1–4, and solving orders were coded as 0–3 rather than 1–3 for modeling purposes. The estimation method used for the analyses was a Gaussian Hermite approximation to the log-likelihood with 100 points per axis. In addition to the models described here, we also ran all models including two main effects and their interaction, as well as a three main effect model including all two-way interactions but excluding the three-way interaction. All models including interactions failed to show improvement over their corresponding main-effects-only models. Including the three-way interaction also failed to affect the conclusions.

* $p < .05$
** $p < .01$
*** $p < .001$

Table 6 Results and comparisons of models predicting the log odds (logit) of trials being solved with a correct strategy from solving order, problem number, and traversing eye movements on timed solving trials, excluding cases where typical arithmetic strategies were used

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
Factors: Model-derived parameters [95 % confidence intervals]								
Intercept	0.90*** [0.54, 1.26]	0.21 [−0.39, 0.81]	−0.31 [−0.94, 0.31]	0.66** [0.20, 1.13]	−1.37** [−2.34, −0.41]	−0.14 [−0.85, 0.56]	−1.79** [−2.90, −0.69]	−2.02* [−3.90, −0.14]
Solving order		0.55* [0.12, 0.98]			0.81** [0.27, 1.36]	0.61** [0.17, 1.04]	0.88** [0.33, 1.43]	1.07 [−0.12, 2.26]
Problem number			0.89*** [0.53, 1.25]		0.93*** [0.56, 1.30]		0.94*** [0.56, 1.31]	1.07* [0.16, 1.98]
Traversals				0.28 [−0.09, 0.66]		0.34 [−0.03, 0.71]	0.38 [−0.04, 0.80]	0.37 [−1.00, 1.74]
Solving order × problem number								−0.13 [−0.77, 0.50]
Solving order × traversals								−0.02 [−0.99, 0.94]
Problem number × traversals								0.01 [−0.69, 0.72]
Solving order × problem number × traversals								0.02 [−0.56, 0.59]
SD of random effect	0.55	0.26	0.95	0.55	0.67	0.20	0.61	0.63
AIC	165.5	161.6	141.7	165.3	135.5	160.3	134.3	142.0
BIC	171.9	171.2	151.3	174.8	148.3	173.0	150.2	170.7
Log likelihood	−80.75	−77.82	−67.9	−79.7	−63.8	−76.1	−62.2	−62.0
Deviance	161.5	155.6	135.7	159.3	127.5	152.3	124.3	124.0
Model Comparisons								
Compared to 1		$\chi^2(1) = 5.9^*$	$\chi^2(1) = 25.8^{***}$	$\chi^2(1) = 2.2$	$\chi^2(2) = 34.0^{***}$	$\chi^2(2) = 9.2^{**}$	$\chi^2(3) = 37.2^{***}$	$\chi^2(7) = 37.5^{***}$
Compared to 2					$\chi^2(1) = 28.1^{***}$	$\chi^2(1) = 3.4$	$\chi^2(2) = 31.3^{***}$	$\chi^2(6) = 31.6^{**}$
Compared to 3					$\chi^2(1) = 8.2^{**}$		$\chi^2(2) = 11.4^{**}$	$\chi^2(6) = 11.7$
Compared to 4						$\chi^+(1) = 7.0^*$	$\chi^2(2) = 35.0^{***}$	$\chi^2(6) = 35.3^{***}$
Compared to 5							$\chi^2(1) = 3.2$	$\chi^2(5) = 3.5$
Compared to 6							$\chi^2(1) = 28.0^{***}$	$\chi^2(5) = 28.3^{***}$
Compared to 7								$\chi^2(4) = 0.3$

$p < .05$ represents a rejection of the null hypothesis that a factor does not significantly predict responding (top: Parameters) or that the models fit equally well (bottom: Model Comparisons). Problem numbers were coded as 0–3 rather than 1–4, and solving orders were coded as 0–3 rather than 1–3 for modeling purposes. The estimation method used for the analyses was a Gaussian Hermite approximation to the log-likelihood with 100 points per axis. In addition to the models described here, we also ran all models including two main effects and their interaction, as well as a three main effect model including all two-way interactions but excluding the three-way interaction. All models including interactions failed to show improvement over their corresponding main-effects-only models. Including the three-way interaction also failed to affect the conclusions

* $p < .05$

** $p < .01$

*** $p < .001$

setting: The majority of math practice is performed alone, with written responses handed in and not returned for several days. Consequently, these school settings may perpetuate the fixation on typical arithmetic strategies. Repeated use of these strategies in the context of math equivalence problems may then compound the issue by reinforcing the association of activated operational schemata to inappropriate situations.

However, feedback is not a panacea, particularly when given without explaining the error, as is typical of markings on students' schoolwork. Consider, for example, water jar problems for which the set strategy is ineffective. In these cases, failure to solve the problem provides immediate feedback that the current strategy is not working. One can immediately determine that the solution strategy $B - 2C - A$ is incorrect by its failure to yield the solution of 25 quarts given 28-, 76-, and 3-quart jars, yet some participants forge ahead with the set strategy or simply give up because they fail to generate alternative strategies (Luchins, 1942). This type of Einstellung effect is unlikely to underlie typical arithmetic strategy use in undergraduates who possess relational schemata. However, it likely plays a role in children's use of typical arithmetic strategies to solve math equivalence problems. Children often continue to use such strategies, even after receiving feedback that they are solving problems incorrectly (Alibali, 1999). Many even revert back to using typical arithmetic strategies after being taught a correct strategy for solving the problems (Alibali, 1999; Cook et al., 2008). This highlights the need for schools to implement curricula that foster the development of relational schemata before operational patterns become entrenched.

Future directions

Although our hypotheses were supported, several questions remain. First, are traversing eye movements simply indicators of the activation of relational schemata, or are they a mechanism of change themselves? Could moving one's eyes between sides of the equal sign actually help activate relational schemata and increase correct strategy use? There is reason to suspect that it would. After first determining what items successful solvers typically fixated on in Duncker's (1945) radiation problem, Grant and Spivey (2003) found that drawing solvers' attention to these items improved solving success. Furthermore, when Thomas and Lleras (2007) guided participants' eye movements on the radiation problem via a tracking task, they found that inducing eye movements that embodied correct solutions improved successful solving rates.

A second, related question is whether inducing traversing eye movements would increase children's learning of relational concepts. Again, there is reason to suspect that it would. As was mentioned above, having children mimic relational gestures supports learning of mathematical equivalence (Cook

et al., 2008; Goldin-Meadow, Cook, & Mitchell, 2009). The back-and-forth eye movements that almost certainly accompany these relational gestures may play a role in the effectiveness of this manipulation.

Conclusions

Our findings support the claim that acquiring operational schemata negatively affects mathematical performance far beyond elementary school. Once acquired, operational schemata are never fully unlearned (McNeil et al., 2010) and can be activated instead of correct relational schemata, yielding long-term Einstellung effects. Fortunately, evidence suggests that even minor modifications to elementary school instructional practices, such as linking the equal sign to inequality symbols (Hattikudur & Alibali, 2010), presenting math equivalence problems with concrete materials (Sherman & Bisanz, 2009), using nontraditional, $c = a + b$ formats (McNeil, Fyfe, Petersen, Dunwiddie, & Brletic-Shipley, 2011), and practicing addition facts organized by equivalent sums (McNeil et al., 2012), support relational schemata development. Although such evidence has not yet persuaded most publishers to include these modifications in their mathematics curricula, studies like this one demonstrating long-term negative consequences of acquiring operational schemata add to a growing chorus pressing for change.

Appendix

Timed Solving Trials

1. $7 + 9 + 6 = 8 +$ __
2. $6 + 8 + 4 = 7 +$ __
3. $9 + 7 + 8 = 5 +$ __
4. $4 + 7 + 8 = 4 +$ __

Untimed Solving Trials

1. $6 + 3 + 9 = 4 +$ __
2. $8 + 5 + 6 = 3 +$ __
3. $7 + 4 + 6 = 5 +$ __
4. $4 + 9 + 3 = 4 +$ __

Encoding Trials

1. $4 + 5 + 7 = 3 +$ __
2. $7 + 5 + 4 = 6 +$ __
3. $5 + 3 + 6 = 4 +$ __
4. $4 + 3 + 5 = 4 +$ __

References

Alibali, M. W. (1999). How children change their minds: Strategy change can be gradual or abrupt. *Developmental Psychology, 35,* 127–145. doi:10.1037/0012-1649.35.1.127

Allport, G. W. (1954). *The nature of prejudice.* Cambridge, MA: Addison-Wesley.

Baroody, A. J., & Ginsburg, H. P. (1983). The effects of instruction on children's understanding of the "equals" sign. *The Elementary School Journal, 84,* 199–212. doi:10.1086/461356

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4 (Version .999375-42) [computer software and manual]: Linear Mixed-effects Models Using S4 Classes.* Accessible from http://cran.r-project.org/web/packages/lme4.html

Bilalić, M., McLeod, P., & Gobet, F. (2008a). Inflexibility of experts - Reality or myth? Quantifying the Einstellung effect in chess masters. *Cognitive Psychology, 56,* 73–102. doi:10.1016/j.cogpsych.2007.02.001

Bilalić, M., McLeod, P., & Gobet, F. (2008b). Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition, 108,* 652–661. doi:10.1016/j.cognition.2008.05.005

Clement, J., Lochhead, J., & Monk, G. (1981). Translation difficulties in learning mathematics. *The American Mathematical Monthly, 88*(4), 286–290.

Cook, S. W., Mitchell, Z., & Goldin-Meadow, S. (2008). Gesture makes learning last. *Cognition, 106,* 1047–1058. doi:10.1016/j.cognition.2007.04.010

Croskerry, P. (2003). The importance of cognitive errors in diagnosis and strategies to minimize them. *Academic Medicine, 78,* 775–780.

Diamond, A., & Kirkham, N. Z. (2005). Not quite as grown-up as we like to think: Parallels between cognition in childhood and adulthood. *Psychological Science, 16,* 291–297. doi:10.1111/j.0956-7976.2005.01530.x

Duncker, K. (1945). On problem solving. *Psychological Monographs, 58* (Whole No. 270).

Ellis, J. J., Glaholt, M. G., & Reingold, E. M. (2011). Eye movements real solution knowledge prior to insight. *Consciousness and Cognition, 20,* 768–776. doi:10.1016/j.concog.2010.12.007

Goldin-Meadow, S., Alibali, M. W., & Church, R. B. (1993). Transitions in concept acquisition: Using the hand to read the mind. *Psychological Review, 100,* 279–297. doi:10.1037/0033-295X.100.2.279

Goldin-Meadow, S., Cook, S. W., & Mitchell, Z. A. (2009). Gesturing gives children new ideas about math. *Psychological Science, 20*(3), 267–272. doi:10.1111/j.1467-9280.2009.02297.x

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science, 14*(5), 462–466. doi:10.1111/1467-9280.02454

Grove, A. (2012). *University of Notre Dame - SAT Scores, Costs and Admissions Data.* Retrieved July 19, 2012, from http://collegeapps.about.com/od/collegeprofiles/p/notre-dame.htm

Hattikudur, S., & Alibali, M. W. (2010). Learning about the equal sign: Does comparing with inequality symbols help? *Journal of Experimental Child Psychology, 107,* 15–30. doi:10.1016/j.jecp.2010.03.004

Jacobs, V. R., Franke, M. L., Carpenter, T. P., Levi, L., & Battey, D. (2007). Professional development focused on children's algebraic reasoning in elementary school. *Journal for Research in Mathematics Education, 38,* 258–288.

James, W. (1890). *The principles of psychology.* New York: Holt.

Kelley, K. (2007a). Methods for the behavioral, educational, and social science: An R package. *Behavior Research Methods, 39*(4), 979–984. doi:10.3758/BF03192966

Kelley, K. (2007b). Confidence intervals for standardized effect sizes: Theory, application, and implementation. *Journal of Statistical Software, 20*(8), 1–24.

Knoblich, G., Ohlsson, S., Haider, H., & Rhenius, D. (1999). Constraint relaxation and chunk decomposition in insight problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1534–1555. doi:10.1037/0278-7393.25.6.1534

Knoblich, G., Ohlsson, S., & Raney, E. G. (2001). An eye movement study of insight problem solving. *Memory & Cognition, 29,* 1000–1009. doi:10.3758/BF03195762

Li, X., Ding, M., Capraro, M. M., & Capraro, R. M. (2008). Sources of differences in children's understandings of mathematical equality: Comparative analysis of teacher guides and student texts in China and the United States. *Cognition and Instruction, 26,* 195–217. doi:10.1080/07370000801980845

Lippman, L. G. (1994). Rule governed performance and sensitivity to contingencies: What's new? *The Journal of General Psychology, 121,* 353–360. doi:10.1080/00221309.1994.9921209

Lovett, M. C., & Anderson, J. R. (1996). History of success and current context in problem solving: Combined influences on operator selection. *Cognitive Psychology, 31,* 168–217. doi:10.1006/cogp.1996.0016

Luchins, A. S. (1942). Mechanization in problem solving – The effect of Einstellung. *Psychological Monographs, 54*(6), i-95.

Maier, N. R. F. (1931). Reasoning in humans: II. The solution of a problem and its appearance in consciousness. *Journal of Comparative Psychology, 12,* 181–194. doi:10.1037/h0071361

McNeil, N. M. (2007). U-shaped development in math: 7-year-olds outperform 9-year-olds on equivalence problems. *Developmental Psychology, 43,* 687–695. doi:10.1037/0012-1649.43.3.687

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,* 451–466. doi:10.1207/s15516709cog2803_7

McNeil, N. M., & Alibali, M. W. (2005a). Knowledge change as a function of mathematics experience: All contexts are not created equal. *Journal of Cognition and Development, 6,* 285–306. doi:10.1207/s15327647jcd0602_6

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

McNeil, N. M., Chesney, D. L., Matthews, P. G., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Wheeler, M. C. (2012). It pays to be organized: Organizing addition knowledge around equivalent values facilitates understanding of mathematical equivalence. *Journal of Educational Psychology.*

McNeil, N. M., Fyfe, E. R., Petersen, L. A., Dunwiddie, A. E., & Brletic-Shipley, H. (2011). Benefits of practicing 4 = 2 + 2: Nontraditional problem formats facilitate children's understanding of mathematical equivalence. *Child Development, 82*(5), 1620–1633. doi:10.1111/j.1467-8624.2011.01622.x

McNeil, N. M., Rittle-Johnson, B., Hattikudur, S., & Petersen, L. A. (2010). Continuity in representation between children and adults: Arithmetic knowledge hinders undergraduates' algebraic problem solving. *Journal of Cognition and Development, 11*(4), 437–457. doi:10.1080/15248372.2010.516421

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the AB task. *Developmental Science, 1*(2), 161–184. doi:10.1111/1467-7687.00021

Ohlsson, S. (2009). Resubsumption: A possible mechanism for conceptual change and belief revision. *Educational Psychologist, 44,* 20–40. doi:10.1080/00461520802616267

R Development Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: F Foundation for Statistical Computing.

Rabe-Hesketh, S., & Skrondal, A. (2012). *Multilevel and longitudinal modeling using Stata* (3rd ed.). College Station, TX: Stata Press.

Rescorla, R. A. (1996). Preservation of pavlovian associations through extinction. *The Quarterly Journal of Experimental Psychology. B, 49*(3), 245–258. doi:10.1080/713932629

Schauble, L. (1990). Belief revision in children: The role of prior knowledge and strategies for generating evidence. *Journal of Experimental Child Psychology, 49*(1), 31–57. doi:10.1016/0022-0965(90)90048-D

Schultz, P. W., & Searleman, A. (2002). Rigidity of thought and behavior: 100 years of research. *Genetic, Social, and General Psychology Monographs, 128*(2), 165–207. doi:10.1024/1421-0185/a000070

Sherman, J., & Bisanz, J. (2009). Equivalence in symbolic and non-symbolic contexts: Benefits of solving problems with manipulatives. *Journal of Educational Psychology, 101,* 88–100. doi:10.1037/a0013156

Stephen, D. G., Boncoddo, R. A., Magnuson, J. S., & Dixon, J. A. (2009). The dynamics of insight: Mathematical discovery as a phase transition. *Memory & Cognition, 37*(8), 1132–1149. doi:10.3758/MC.37.8.1132

Thomas, L. E., & Lleras, A. (2007). Moving eyes and moving thought: On spatial compatibility between eye movements and cognition. *Psychonomic Bulletin & Review, 14*(4), 663–668. doi:10.3758/BF03196818

Wiley, J. (1998). Expertise as mental set: The effects of domain knowledge in creative problem solving. *Memory & Cognition, 26*(4), 716–730. doi:10.3758/BF03211392

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development, 11*(1), 37–63. doi:10.1016/S0885-2014(96)90027-1