American Psychological Association

Journal of Educational Psychology

© 2018 American Psychological Association
0022-0663/19/$12.00

2019, Vol. 111, No. 3, 414–433
http://dx.doi.org/10.1037/edu0000294

Individual Differences in Addition Strategy Choice: A Psychometric Evaluation

Katherine T. Rhodes
The Ohio State University

Sarah Lukowski
University of Minnesota

Lee Branum-Martin
Georgia State University

John Opfer
The Ohio State University

David C. Geary
University of Missouri

Stephen A. Petrill
The Ohio State University

The strategy choice model (SCM) is a highly influential theory of human problem-solving. One strength of this theory is the allowance for both item and person variance to contribute to problem-solving outcomes, but this central tenet of the model has not been empirically tested. Explanatory item response theory (EIRT) provides an ideal approach to testing this core feature of SCM, as it allows for simultaneous estimation of both item and person effects on problem-solving outcomes. We used EIRT to test and confirm this central tenet of the SCM for adolescents' ($n = 376$) solving of addition problems. The approach also allowed us to identify the strategy choices of adolescents who still struggle with basic arithmetic. The synthesis of SCM theory and EIRT modeling has implications for more fully investigating the sources of individual differences in students' problem solving, and for identifying problem-solving patterns associated with poor academic achievement.

Educational Impact and Implications Statement

Strategy usage may be an important avenue for the identification and treatment of mathematics difficulties. The current study examined individual differences in strategic problem-solving behavior with a sample of adolescent problem-solvers. Results indicate that even in adolescence there are meaningful individual differences in how students solve addition problems, including some adolescents who continue to rely on immature counting strategies. Those adolescents who still relied on immature counting strategies were also struggling with broad mathematical achievement. These struggling students can be identified with a brief addition strategy assessment, and our results suggest that they may benefit from mastering more developmentally mature strategies. For students with mathematical difficulties, simply advancing to developmentally mature strategy selection may be an important intervention goal, one which is often overlooked in educational settings.

Keywords: strategy choice model, explanatory item response theory, arithmetic problem-solving, individual differences, adolescent cognitive addition

This article was published Online First September 27, 2018.

Katherine T. Rhodes, Department of Psychology, The Ohio State University; Sarah Lukowski, Institute of Child Development, University of Minnesota; Lee Branum-Martin, Department of Psychology, Georgia State University; John Opfer, Department of Psychology, The Ohio State University; David C. Geary, Department of Psychological Sciences, University of Missouri; Stephen A. Petrill, Department of Psychology, The Ohio State University.

We thank participating families and adolescents as well as the support staff across participating research sites. Katherine T. Rhodes would like to extend a special thank you to Paul De Boeck for his statistical advice in preparing this article. The authors also would like to thank Maggie Beard, Renee Brooks, and Alicia Nunez for their tireless efforts in data management for the current study. Study data were collected and managed using REDCap electronic data capture tools hosted at The Ohio State University (see Harris et al., 2009). This research was completed in part under funding from the Institute of Education Sciences (IES) under award R305A160295 and from the Eunice Kennedy Shriver National Institute of Child Health and Human Development of the National Institutes of Health under award HD095215, HD075460, and HD090868. The content is solely the responsibility of the authors and does not necessarily represent the official views of the IES or the National Institutes of Health.

Correspondence concerning this article should be addressed to Katherine T. Rhodes, Department of Psychology, The Ohio State University, 1835 Neil Avenue, Columbus, OH 43210. E-mail: rhodes.390@osu.edu

Across the life span, the strategies used during problem-solving are an important indicator of mathematical cognition. Since its original formulation over 30 years ago, the strategy choice model (SCM) has been one of the most influential conceptualizations of the development of problem-solving competencies and is well characterized for arithmetical problem solving (Bailey, Littlefield, & Geary, 2012; Geary, Widaman, Little, & Cormier, 1987; Geary & Wiley, 1991; Geary, Hoard, Byrd-Craven, & DeSoto, 2004; Siegler, 1986, 1987a, 1988a; Siegler & Robinson, 1982; Siegler & Shrager, 1984). A core assumption of the SCM—that problem solving is a joint function of item difficulty and individual-level knowledge—was assessed in the early phases of theory development (Siegler, 1987b) but has not yet been evaluated with more recent and nuanced analytic methods that allow both items and persons to differ in their contributions to problem-solving outcomes. This omission means that important individual differences between problems and problem-solvers are potentially obscured in favor of reporting average trends. Accordingly, we evaluate this core assumption of the SCM using an explanatory item response theory (EIRT) modeling method; specifically, allowing for individual differences or "random effects" for both items and persons. The method was applied to a sample of adolescent problem-solvers who are understudied in the context of the SCM literature. We evaluate the utility of the method for identifying struggling adolescent problem-solvers based on the strategies used to solve arithmetic problems and demonstrate convergent and discriminant validity of the strategy parameters using measures of broad achievement in mathematics and reading. The individual differences between adolescent problem-solvers would not have been distinguishable using the analytic methods traditionally used in evaluating the SCM.

The SCM

The SCM outlines the processes underlying people's use of one problem-solving approach or another to solve any particular problem as well as the mechanisms that govern developmental change in the mixture of strategies used during problem solving (Kerkman & Siegler, 1993, 1997; Siegler, 1986, 1987b, 1988a, 1991, 1996; Siegler & Shrager, 1984; Siegler & Taraban, 1986). The central tenet is that people's adaptive strategy choices require balancing conflicting problem-solving goals (e.g., speed and accuracy) and managing problem-specific demands (e.g., varying levels of problem difficulty) and cognitive constraints (e.g., the ability to retrieve a solution from memory). Arguably, it is the incorporation of variation in problem demands, individual differences in domain-specific knowledge, and human strategic behaviors that has contributed to the SCM's continued influence in cognitive, developmental, and educational psychology.

Siegler's initial formulation of the SCM was to explain children's solving of addition and subtraction problems, and this formulation subsequently proved successful in explaining strategy choices in multiplication, spelling, balance scales, reading, and telling time, among other domains (Siegler, 1986, 1987a, 1988b, 1991, 1996; Siegler & Jenkins, 1989; Siegler & McGilly, 1989; Siegler & Shrager, 1984). With respect to arithmetic, the SCM has been successfully used to understand developmental changes in the strategy mix (Siegler & Jenkins, 1989), cross-cultural differences in children's strategy choices and cross-generational differences in adults' choices (Geary, Fan, & Bow-Thomas, 1992; Geary, Frensch, & Wiley, 1993), as well as the problem-solving patterns and underlying cognitive deficits that are common in children with mathematical learning difficulties (MLD; Geary & Brown, 1991). As a result, addition is perhaps the best understood domain of strategy development.

Typical and Atypical Addition Strategy Development

At school entry, most children use a combination of finger counting and verbal counting to solve the majority of addition problems (Siegler & Shrager, 1984). With the former, children lift their fingers to physically represent the addends and then count them to reach a sum. During verbal counting, children count audibly or move their lips as if counting implicitly. Whether or not they use their fingers, children sometimes count both addends starting from 1 (sum strategy), start with the smaller addend and count the larger one (max strategy), or start with the larger addend and count the smaller one (min strategy). A critical prediction of the SCM is that the use of counting results in the formation of an associative relation between the problem and the generated answer.

Once associative memories between problem addends and the answer generated by counting are formed, the child will begin to use retrieval in problem solving. So, when the problem is presented again, memory representations of counting schema and the answer stored in long-term memory compete for expression. If the activation level of the counting schema exceeds the strength of the problem–answer association (associative strength), then the child will count to solve the problem. Repeated use of counting builds this associative strength and eventually results in consistent use of retrieval-based problem solving (Siegler, 1996; Siegler & Shrager, 1984). Decomposition is one common retrieval-based strategy (Siegler, 1987c). The problem 6 + 8 might be solved by decomposing 8 into 4 and 4, then retrieving the answer to 6 + 4, and finally adding back the other 4. The use of decomposition is also dependent on a conceptual understanding of number relations (Geary et al., 2004). With sufficient practice, most children will directly retrieve the answer from long-term memory to solve most simple problems, although many students in the United States do not receive this level of practice and thus continue using a mix of counting and retrieval-based strategies into adulthood (Geary & Wiley, 1991).

On top of amount of practice, individual differences in children's strategy choices are related to their confidence in the accuracy of the retrieved answer (Siegler, 1988a). "Perfectionists" will often resort to min-counting to verify the accuracy of retrieved answers, whereas "not-so-good" students state any answer that comes to mind, whether or not it is likely to be correct. There also appear to be important individual differences in the hippocampal-dependent memory system that underlies the formation of problem-answer associative memories (Qin et al., 2014), and this system has been implicated in the retrieval deficits that are a cardinal feature of long-term learning difficulties in mathematics (Geary, 1993; Supekar et al., 2013).

These retrieval deficits have been well documented in elementary schoolchildren (e.g., Geary & Brown, 1991). These children can retrieve the answer to some basic addition problems, but to solve other problems they typically have to resort to min or sum-counting. The persistence of these deficits has been confirmed in longitudinal studies of elementary schoolchildren (Geary, Hoard, & Bailey, 2012; Jordan, Hanich, & Kaplan, 2003) and predicts later difficulties memorizing the basic structure of algebra equations (Geary, Hoard, Nugent, & Rouder, 2015). In the latter study, earlier retrieval deficits predicted later algebra difficulties in high school, but addition strategy choices were not directly assessed in high school. In fact, little is known about the retrieval deficits and continued reliance on counting strategies in adolescents who have difficulties with mathematics learning. The current study addresses this gap in the literature.

Empirical Evaluation of the Strategy Choice Model

Numerous studies have confirmed the basic processes identified in the SCM (Kerkman & Siegler, 1993, 1997; Siegler, 1986, 1987b, 1988a, 1991, 1996; Siegler & Shrager, 1984; Siegler & Taraban, 1986). The evaluation of these processes is based on differences between strategies, solution times, and accuracies. For example, associative strengths between problems with smaller addends (e.g., 2 + 3) and their sums are predicted to form earlier than for problems with larger addends (e.g., 8 + 7). Empirically, retrieval should be used more frequently to solve problems with smaller addends and the accompanying solution times should be faster and accuracies higher than for problems with larger addends (Siegler, 1987c). In other words, strategy frequencies, solution times, and accuracy percentages vary systematically with problem difficulty. Strategies also differ in systematic ways, whereby retrieval and decomposition are executed more quickly than counting strategies and min-counting is generally more accurate than sum-counting due to fewer counts and thus fewer opportunities to commit an error (Geary et al., 2012; Siegler, 1987c; Siegler & Shrager, 1984).

Most SCM analyses to date have focused on describing strategies and comparing their effects on problem-solving outcomes—how often types of strategies are used, how effective different strategies are in quickly and accurately achieving solutions, when different types of strategies emerge during development, and the extent to which strategy usage can be employed to identify problem-solvers with different cognitive profiles. For example, calculating the relative frequency of a strategy entails tallying its use across a set of problems and then calculating an average ratio across a sample of problem-solvers who have attempted those problems. Similarly, calculating the relative efficiency of a strategy entails averaging the accuracy and solution time on problems in which it was used across a sample of problem-solvers. These averages can then be compared between problem-solvers of a particular age, problem-solvers of different cognitive profiles, and even across problems of a certain type (e.g., single digit items vs. double digit items). Thus, the methods used to evaluate the SCM to date have necessarily relied on three main analytic methods: (a) mean differences, (b) bivariate correlations, and (c) multiple regressions. However, these traditional methods all rely either on collapsing across repeated items to assess differences across people (as described above) or on collapsing across persons to analyze differences across problems.

Collapsing across problems or persons requires the assumption of negligible intraclass variability, an assumption that is often unfounded because level members (e.g., items or persons) frequently do vary substantially from one another (Embretson, 1983). Indeed, in practice as much as 80% to 90% of the variance in outcomes may be accounted for by "random" differences between members of a within-group cluster (e.g., children nested in a school, schools nested in a district; Raudenbush & Bryk, 2002). In these cases, an important source of variance in the data is simply relegated to "error" terms and otherwise hidden in averaged results. EIRT is a modern psychometric method capable of avoiding this confound by simultaneously estimating problem and person effects and, in doing so, provides a more complete and nuanced assessment of the SCM than is possible with traditional analytic methods.

Evaluating the SCM With EIRT

EIRT combines experimental design (the goal being to explain a dependent variable in terms of the experimental design factors) and observational measurement (the goal being to estimate individuals' traits on a construct or set of constructs; De Boeck & Wilson, 2004). Thus, the goal of EIRT is explanatory measurement, to provide measurement models capable of both describing individual traits and explaining individual differences (De Boeck & Wilson, 2004). Essentially, EIRT is a special name given to models of multilevel, categorical responses, in which latent estimates are generated for both item and person parameters, and explanatory variables can be modeled as predictors of responses across levels of analysis (Baayen, Davidson, & Bates, 2008; De Boeck & Wilson, 2004). EIRT is especially well-suited for examining theories such as the SCM.

In general, models used in EIRT fall within the broad family of generalized linear mixed models and nonlinear mixed models that are traditionally used to analyze item responses. However, EIRT models may vary with respect to their specifications about (a) the modeled data structures (i.e., the model predictors as item-level, person-level, or cross-classified—meaning that units of analysis are simultaneously nested under two levels, such as children nested in both neighborhoods and schools), and (b) the nature of model effects (i.e., fixed vs. random effects, where fixed effects, represent the average outcome for all group members, with random components, represented by allowing each individual to deviate from this average). Tailoring an EIRT model to correctly evaluate the SCM requires mapping central theoretical postulates of the SCM onto statistical specifications inherent in an EIRT model.

Evaluating the SCM for individual differences should incorporate the theoretical ideas that people engage in adaptive problem solving by (a) using a variety of strategies across a variety of problems, (b) managing a variety of problem demands and cognitive constraints, and (c) optimizing the potential trade-offs of problem solving (e.g., speed vs. accuracy). In statistical terms, the first specification can be met with a cross-classified model, such that adaptive problem-solving responses are nested in both persons and items (and strategy usage is a predictor of these responses at both person and item-levels; (Hill & Goldstein, 1998; Matuschek, Kliegl, Vasishth, Baayen, & Bates, 2017). In particular, the effect of strategy use may be different for different persons (i.e., some may use it better than others). The second specification can be met with modeling mixed effects, both fixed and random effects for items and persons, such that across items and persons an average effect can be observed but individual differences are allowed (for a more in-depth review of mixed effect modeling, see, e.g., Raudenbush & Bryk, 2002). The third specification can be met with a bivariate model, at the very least, in which both accuracy and solution time are simultaneously accounted for as outcomes (for more detail on the SCM specifications for relations between solution time and accuracy, see, e.g., Siegler, 1991, 1996; Siegler & Robinson, 1982; Siegler & Shrager, 1984; Siegler & Taraban, 1986).

The Current Study

In the current study, we used an EIRT framework to provide the first simultaneous evaluation of both problem and person effects predicated by the SCM. We focused on problem solving in addition because variation in the associated strategy choices are well characterized across problems and individuals. Moreover, we focused on adolescents' problem solving, which is surprisingly understudied. More practically, we evaluated the EIRT model for its capacity to (a) identify struggling adolescent problem-solvers and (b) demonstrate convergent and discriminant validity with measures of broad mathematics and reading achievement.

Method

Participants

Participants were 376 (224 female) adolescents (M = 15.05, SD = 1.45, range = 11 to 18 years) drawn from a longitudinal twin study conducted in the Midwestern United States, which was approved by The Office of Responsible Research Practices at The Ohio State University (Petrill, Deater-Deckard, Thompson, DeThorne, & Schatschneider, 2006). The parent study is examining the genetic and environmental predictors of reading and mathematics skill development. Participants were initially recruited largely via school nominations of families with twins in kindergarten or 1st grade. Throughout the Greater Cleveland, Columbus, and Cincinnati metropolitan areas in Ohio and in Western Pennsylvania, participating schools (n = 273) forwarded study information to eligible families. Additional families were recruited via birth records, twin family clubs, and media advertisement. Those families who expressed interest in the study were contacted by telephone and, if interested in participation, mailed a consent letter and demographic questionnaire requesting demographic information. Once consent and the initial demographic survey were returned, participants assented and were assessed in their homes approximately once annually over 10 waves.

The data for these analyses come from the ninth measurement, which was one of three waves focused intensively on mathematical skill development. Wave 9 included measures of arithmetic strategy usage and its associated cognitive predictors. The sample for the current study was largely White (90%) and middle-class (the majority of participating families were in two-parent households in which both parents had obtained 4-year college degrees). Though participants ranged from 5th to 12th grade, most were in middle-school (65% were between 7th and 10th grades), and few had identified disabilities (7%; mostly related to ADHD and/or learning difficulties). Participant demographics are displayed in Table 1.

Table 1
Participant Demographics on Categorical Variables

Variable and category	Frequency	Percentage
Gender		
Male	152	40
Female	224	60
Grade		
Not answered	54	14
5	4	1
6	22	6
7	32	9
8	80	21
9	90	24
10	42	11
11	40	11
12	12	3
Race		
Not answered	4	1
Asian/Asian American	8	2
African American/Black	18	5
Hispanic	2	1
European American/White	338	90
Other	6	2
Current IEP		
Not answered	120	32
No	229	61
Yes	27	7

Note. IEP = individual education plan.

Measures

Addition strategy. Thirty addition items that varied in difficulty (based on addend size) were created based on Geary and colleagues' (2004) addition strategy assessment. Participants were shown addition problems one at a time on a computer. Addition Items 1 through 14 contained two single digit addends (e.g., 3 + 6); Addition Items 15 through 20 contained one single digit addend and one double digit addend (e.g., 16 + 7); and Addition Items 21 through 30 contained two double digit addends (e.g., 13 + 24). All participants were shown the same items in the same order.

Problems were presented one at a time on the center of a computer screen. Participants were asked to solve each problem as quickly and accurately as possible. When the participant said the answer aloud, the examiner pressed the space bar on the computer to measure the solution time. Prior work has suggested that interviewing is both a valid and efficient way to obtain information about strategy usage (Siegler, 1987b), and so once a participant answered the problem they were asked "How did you figure out the answer to that problem?" Participant responses were audio recorded. Two trained research assistants independently reviewed audio recordings and coded participants' strategy responses into several applicable categories using Geary and colleagues' (2004) coding scheme. The independently rated strategy codes were then compared for agreement. In the event of coding disagreements, the two raters met, reviewed the audio recordings, and discussed the coding scheme until consensus could be reached. In the event that this consensus meeting still resulted in a coding disagreement, a senior researcher met with both raters to discuss codes and arrive at consensus. Thus all strategy responses in the current study were the result of 100% agreement between a minimum of two independent raters.

On items in which there were two single digit addends or one single digit addend and a double-digit addend, participants reported using a counting (either with their fingers or verbally), decomposition (breaking the problem into smaller parts), or retrieval strategy (remembering the answer). The counting strategies were further categorized based on whether the participant reported counting both addends (sum), the largest addend (max), or the smallest addend (min). For the double digit problems, participants typically reported using an algorithm (adding the ones and then adding the tens), in keeping with process models of complex arithmetic (Geary et al., 1987; Widaman, Geary, Cormier, & Little, 1989). Thus, the task results in three features of problem solving: solution times (s), accuracy (correct = 1, incorrect = 0), and self-reported strategy. Descriptive statistics for participants' accuracies and solution times can be found in Table 2, and Table 3 presents accuracies and solution times by participants' most frequently used strategies. The pattern of solution times, with retrieval being the fastest, followed by decomposition and counting are consistent with previous studies (Geary et al., 2012; Siegler, 1987b) and suggests the self-reports provided a valid measure of how the problems were solved (see the Results section).

Mathematical ability. Mathematics ability was assessed with three subtests of the Woodcock-Johnson III Tests of Achievement (McGrew & Woodcock, 2001). The Calculation subtest assessed children's ability to computationally solve a variety of mathematics problems. These included addition, subtraction, multiplication, and division computations, along with problems that included combinations of these operations. Computations included negative numbers, percentages, decimals, fractions, and whole numbers. Published median internal consistency reliability for this test is .85 in the 5- to 19-year-old age range; for this sample, α = .84. The Applied Problems subtest assessed ability on math story problems. Children completed a series of story problems presented visually and read aloud, requiring them to use a variety of math operations in order to solve questions of increasing difficulty. The published median reliability of this task is .92 in the 5- to 19-year-old age range; for this sample, α = .83. The Math Fluency subtest assessed ability to solve single digit addition, subtraction, and multiplication items in a timed setting. Participants had 3 min to complete as many problems as they could out of a set of 160 items. The published median reliability for this test is .89 for the 7- to 19-year-old age range; for this sample, α = .88. Using Compuscore software, standard scores for each math subtest were generated for each subtest, and the scores from these three subtests were combined into a broad math composite score.

Reading ability. Reading ability was assessed with the two subtests of the Test of Word Reading Efficiency (Torgesen, Wagner, & Rashotte, 1999). In Sight Word Efficiency, participants read a list of real words printed in vertical lists. Similarly, in Phonemic Decoding Efficiency, participants read a list of pronounceable

Table 2
Item Descriptives for Accuracy and Solution Time Across All Strategies Used

		Accuracy		Solution time			Correlations		
Item no.	Content	*M*	*SD*	*M*	*SD*	Range	Acc. & Total	Sol. Time & Total	Acc. & Sol. Time
1	3 + 6	.99	.09	1.65	.61	.850–4.988	.06	−.23***	.04
2	5 + 3	.98	.13	1.64	.81	.828–8.746	.24***	−.24***	−.06
3	7 + 6	.94	.24	3.04	1.71	.987–13.703	.19***	−.19***	−.01
4	3 + 5	1.00	0	1.58	1.05	.791–14.224	—	−.34***	—
5	8 + 4	.97	.16	1.98	1.01	.808–6.007	.05	−.20***	−.14*
6	2 + 8	.99	.09	1.49	.56	.628–5.452	.02	−.28***	.002
7	9 + 7	.97	.17	2.71	1.90	.962–18.331	.24***	−.16**	−.16**
8	2 + 4	.98	.15	1.58	.52	.757–3.884	.16**	−.25***	−.06
9	9 + 5	.96	.20	2.46	3.22	1.028–51.884	.33***	−.19***	−.12*
10	7 + 2	.97	.16	1.69	.69	.851–6.233	.19***	−.19***	−.04
11	9 + 8	.97	.17	2.62	1.77	1.006–14.032	.35***	−.25***	−.20***
12	4 + 7	.93	.25	2.61	1.31	.904–8.242	.28***	−.16**	−.06
13	2 + 5	.98	.14	1.72	.75	.869–7.274	.17**	−.19***	−.07
14	3 + 9	.98	.13	2.07	.83	.977–6.274	.04	−.19***	.04
15	16 + 7	.90	.31	5.47	4.06	1.220–53.355	.33***	−.07	−.09
16	3 + 18	.96	.19	2.58	1.19	1.076–9.545	.31***	−.15**	−.12*
17	9 + 15	.94	.23	4.69	2.99	1.293–28.387	.31***	−.20***	−.04
18	17 + 4	.97	.18	3.06	1.91	.570–18.675	.20***	−.18**	.003
19	6 + 19	.93	.26	4.08	3.32	1.391–42.894	.38***	−.16**	−.08
20	14 + 8	.92	.26	4.70	2.99	1.199–21.481	.17***	−.22***	−.13*
21	13 + 24	.93	.25	5.39	3.38	1.637–34.318	.34***	−.24***	−.09
22	17 + 75	.80	.40	9.67	6.58	2.524–45.589	.46***	−.16**	.03
23	26 + 15	.92	.28	6.79	5.31	1.220–43.374	.44***	−.14*	−.07
24	13 + 17	.92	.27	4.44	4.80	1.227–74.031	.35***	−.15**	−.11*
25	62 + 27	.90	.30	6.36	4.64	1.826–38.406	.39***	−.34***	−.18***
26	25 + 33	.93	.26	5.47	3.96	1.725–35.335	.33***	−.33***	−.20***
27	24 + 18	.89	.32	7.90	6.84	1.357–52.640	.45***	−.12*	−.23***
28	23 + 38	.90	.29	7.62	6.04	2.121–89.657	.34***	−.22***	−.14*
29	65 + 28	.83	.37	10.78	10.17	2.369–129.736	.51***	−.18***	−.13*
30	38 + 36	.88	.33	9.17	7.75	2.693–117.222	.30***	−.13*	−.06

Note. Acc. = accuracy; Sol. = solution.
$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .001$.

Table 3
Item Descriptives for Accuracy and Solution Time for Most Frequent Strategies Used

Item		Frequency strategies		Solution time				Accuracy		
No.	Content	*n*	Strategy	*n*	*M*	*SD*	Range	*n*	*M*	*SD*
1	3 + 6	277	Retrieve	257	1.505	.422	.850–3.427	277	.993	.085
		67	Min count	62	2.204	.913	.984–4.988	67	.985	.122
		9	Count other	7	1.717	.290	1.295–2.044	9	1.000	0
2	5 + 3	277	Retrieve	257	1.420	.477	.828–4.924	277	.993	.085
		85	Min count	80	2.341	1.201	.949–8.746	85	.953	.213
		8	Decompose	7	1.471	.326	1.215–2.055	8	1.000	0
3	7 + 6	160	Decompose	150	3.221	1.477	1.292–9.498	160	.962	.191
		125	Retrieve	115	2.029	1.160	.987–9.222	125	.912	.284
		80	Min count	73	4.265	2.015	1.407–13.703	80	.938	.244
4	3 + 5	303	Retrieve	280	1.369	.396	.791–3.483	303	1.000	0
		57	Min count	53	2.622	2.219	1.030–14.224	57	1.000	0
		8	Decompose	8	1.492	.491	1.079–2.403	8	1.000	0
5	8 + 4	180	Retrieve	168	1.514	.519	.808–3.828	180	.983	.128
		89	Min count	84	2.944	1.157	1.220–5.982	89	.966	.181
		79	Decompose	75	1.971	.957	.934–6.007	79	.962	.192
6	2 + 8	308	Retrieve	286	1.383	.427	.628–3.792	308	.990	.098
		56	Min count	51	2.040	.835	.996–5.452	56	1.000	0
		4	Count other	4	1.966	.515	1.232–2.436	4	1.000	0
7	9 + 7	221	Decompose	201	2.441	1.034	1.204–7.065	221	.986	.116
		84	Retrieve	80	1.769	.630	.962–5.297	84	.952	.214
		64	Min count	61	4.523	2.776	1.596–14.625	64	.938	.244
8	2 + 4	296	Retrieve	275	1.487	.428	.757–3.884	296	.986	.116
		52	Min count	49	1.973	.620	1.021–3.857	52	.962	.194
		9	Count other	8	1.678	.776	1.013–3.498	9	1.000	0
9	9 + 5	199	Decompose	180	2.031	1.187	1.133–15.429	199	.990	.100
		101	Retrieve	96	1.651	.487	1.028–4.024	101	.960	.196
		67	Min count	65	4.807	6.652	1.357–51.884	67	.881	.327
10	7 + 2	250	Retrieve	231	1.541	.541	.851–5.390	250	.984	.126
		105	Min count	101	2.037	.867	1.143–6.233	105	.943	.233
		14	Decompose	12	1.782	.614	1.042–3.116	14	1.000	0
11	9 + 8	215	Decompose	196	2.378	.995	1.019–7.331	215	.995	.068
		100	Retrieve	95	1.753	.541	1.006–3.754	100	.980	.141
		48	Min count	46	5.176	3.003	1.506–14.032	48	.833	.377
12	4 + 7	133	Retrieve	123	1.875	.831	.904–6.400	133	.947	.224
		124	Decompose	114	2.663	1.196	1.211–8.242	124	.935	.247
		108	Min count	102	3.383	1.385	1.348–7.768	108	.907	.291
13	2 + 5	278	Retrieve	257	1.573	.622	.869–7.274	278	.986	.119
		87	Min count	81	2.125	.870	1.065–6.945	87	.966	.184
		5	Decompose	5	1.397	.196	1.167–1.666	5	1.000	0
14	3 + 9	145	Decompose	130	2.144	.748	1.068–4.943	145	.979	.143
		137	Retrieve	129	1.706	.564	.977–3.788	137	1.000	0
		70	Min count	68	2.499	.921	1.290–5.476	70	.971	.168
15	16 + 7	221	Decompose	202	5.357	4.755	1.417–53.355	221	.910	.288
		132	Min count	125	5.766	2.717	1.750–20.275	132	.894	.309
		10	Retrieve	10	2.897	2.347	1.220–9.272	10	.900	.316
16	3 + 18	173	Decompose	157	2.478	1.059	1.076–6.059	173	.960	.198
		129	Min count	121	3.091	1.377	1.469–9.545	129	.953	.211
		68	Retrieve	66	1.919	.554	1.154–3.975	68	.985	.121
17	9 + 15	251	Decompose	231	4.147	2.847	1.293–28.387	251	.964	.186
		96	Min count	92	6.179	2.614	2.037–16.023	96	.917	.278
		14	Retrieve	13	2.118	.823	1.387–4.220	14	.786	.426
18	17 + 4	207	Decompose	188	2.905	1.475	1.180–10.124	207	.971	.168
		110	Min count	104	3.811	2.513	.570–18.675	110	.936	.245
		52	Retrieve	50	1.976	1.004	1.131–7.911	52	1.000	0
19	6 + 19	243	Decompose	222	3.840	3.769	1.391–42.894	243	.955	.208
		98	Min count	93	4.990	1.945	1.763–13.176	98	.878	.329
		27	Retrieve	26	2.577	1.913	1.417–11.123	27	.926	.267
20	14 + 8	230	Decompose	212	4.399	2.785	1.199–21.481	230	.930	.255
		105	Min count	98	5.963	3.266	2.125–17.252	105	.895	.308
		28	Retrieve	27	2.247	.971	1.254–5.801	28	1.000	0
21	13 + 24	174	Algorithm	159	5.535	3.073	1.963–20.775	174	.914	.281
		167	Decompose	159	4.564	2.095	1.637–13.291	167	.970	.171
		24	Min count	22	8.202	3.996	3.381–21.868	24	.833	.381
22	17 + 75	185	Decompose	177	8.180	4.836	2.524–37.224	185	.811	.393
		160	Algorithm	144	10.309	6.686	2.628–43.122	160	.800	.401
		19	Min count	17	18.601	10.019	8.602–45.589	19	.737	.452

(*table continues*)

Table 3 (*continued*)

Item		Frequency strategies		Solution time				Accuracy		
No.	Content	*n*	Strategy	*n*	*M*	*SD*	Range	*n*	*M*	*SD*
23	26 + 15	181	Decompose	173	5.785	3.676	1.220–25.688	181	.934	.249
		161	Algorithm	145	7.511	6.147	1.991–43.374	161	.919	.273
		16	Min count	14	11.291	7.994	4.269–37.328	16	.875	.342
24	13 + 17	199	Decompose	189	3.763	2.385	1.263–24.625	199	.930	.256
		122	Algorithm	109	5.004	3.820	1.843–28.582	122	.893	.310
		24	Retrieve	23	2.358	.648	1.227–3.759	24	1.000	0
25	62 + 27	188	Algorithm	170	6.118	4.216	1.926–33.733	188	.899	.302
		163	Decompose	156	5.854	3.678	1.826–31.132	163	.920	.272
		12	Min count	11	15.729	8.869	6.209–38.406	12	.750	.452
26	25 + 33	186	Algorithm	168	5.109	2.928	1.725–20.181	186	.941	.237
		168	Decompose	160	5.277	3.645	1.725–24.063	168	.935	.248
		10	Min count	9	14.771	9.696	4.241–35.335	10	.700	.483
27	24 + 18	187	Decompose	180	6.503	4.611	1.958–45.689	187	.930	.255
		161	Algorithm	146	8.813	7.211	1.357–49.555	161	.863	.345
		14	Min count	13	18.106	15.079	4.336–52.640	14	.714	.469
28	23 + 38	181	Algorithm	162	7.810	4.342	2.308–31.474	181	.912	.285
		176	Decompose	168	6.690	3.451	2.121–19.905	176	.909	.288
		10	Min count	9	21.967	25.786	6.439–89.657	10	.700	.483
29	65 + 28	188	Decompose	178	8.579	4.682	2.369–27.367	188	.872	.335
		164	Algorithm	149	11.793	9.492	2.609–85.214	164	.817	.388
		11	Min count	10	34.864	34.934	7.915–129.736	11	.545	.522
30	38 + 36	190	Decompose	182	7.868	3.972	2.922–30.506	190	.879	.327
		164	Algorithm	147	9.724	5.801	2.693–32.761	164	.902	.298
		6	Min count	6	31.556	42.660	6.018–117.222	6	.833	.408

nonwords. In each subtest, the total number of words correctly read within 45 s was recorded as the raw score. These raw scores were summed and a total word reading efficiency standard score was calculated for each child. Published average test–retest reliability coefficients, across age intervals, for both subtests and the total were .93 to .96; for this sample, α = .75 and .74 for Sight Word Efficiency and Phonemic Decoding, respectively.

Demographic survey. The surveys included questions about home environment, school grades and grade-levels, behavioral problems, and current disability diagnoses and psychological problems, as well as race, income and related information. The questionnaires were collected during home visits. If a family had not had time to complete the survey, an additional survey was mailed with a prestamped envelope. This survey was completed and returned for 254 participants (68%) in the current study.

Procedures

Data collection. Teams of two trained research assistants administered a 3-hr cognitive battery to the twin participants in their homes. These assessments were conducted one-on-one in quiet spaces within the home with allowances for testing breaks as needed. Twins were assessed by separate examiners in separate rooms throughout the visit.

Data analyses. Analyses examined a series of EIRT models in which responses to addition strategy items were nested in both items and persons. However, modeling the nonindependence of twin data was not the focus of the current study, and controlling for this highest level of dyadic clustering was not possible with available software. Therefore, twin pairs were split into random singletons using random number generation in SAS Version 9.3 (SAS Institute Inc., 2011). These two data sets were analyzed independently such that the second set of singletons was used to examine the replicability results of the EIRT models.

Analyses began with an examination of frequencies and descriptive statistics for outcomes across the most frequently used strategies to solve the 30 presented problems; that is, min-counting, decomposition, retrieval, and addition algorithm. Next, a series of EIRT models in which both item and person differences were specified to explain relationships between adolescents' strategy choices and their solution times and accuracy. These EIRT models were examined using Bayesian estimation in Mplus 7 (Muthén & Muthén, 2012). Missing outcome data were estimated using full information maximum likelihood estimation (see e.g., Enders & Bandalos, 2001) in Mplus 7 (Muthén & Muthén, 2012). By default in Mplus 7, missing exogenous predictors were subject to listwise deletion. Finally, EIRT models parameters were examined descriptively for individual differences (in items and persons) and using regression analyses for convergent and discriminant validity (Campbell & Fiske, 1959) with measures of broad academic achievement.

Results

Frequencies and Descriptive Statistics for Outcomes

One would expect to see less variation in adolescents' than children's arithmetic performance. Indeed, there were ceiling effects for accuracy across addition items and problem-solving strategies (mean accuracy ranged from .80 to 1.00; see Table 2). However, consistent with SCM, across both items and persons, there was considerable variance in solution time (solution times ranged from .57 to 129.74 s; see Table 2) and strategy choices (strategy choices ranged from developmentally immature sum-counting to more advanced fact retrieval; see Table 3).

In general, solution times tended to increase as problem sizes increased. Problem-solvers were fastest on single digit items, followed by mixed digit items. They were slowest on double digit items. As expected, the retrieval strategy was the most rapid (M = 1.60 s, SD = .67 s), followed by decomposition (M = 4.54 s, SD = 3.67 s), min-counting (M = 4.70 s, SD = 6.60 s), and addition algorithm strategies (M = 7.76 s, SD = 6.09 s), respectively.

Even among these adolescent problem-solvers, overt counting, especially min-counting, was still common; 20% of simple (e.g., 4 + 6), 30% of mixed (e.g., 17 + 5), and 44% of double (e.g., 26 + 15) problems were solved at least once with min-counting. Some adolescents also reported use of other counting strategies (e.g., fingers and sum-counting), but with relatively low frequency. Across items, the most frequent strategies used were retrieval, decomposition, min-counting, and addition algorithm depending upon item type (i.e., single digit vs. double digit). Retrieval strategies were especially frequent on single digit addition items (54%); addition algorithm strategies were highly frequent on double digit items (44%); decomposition and min strategies were used across all items.

Explanatory Item Response Theory Models

The "build-up" approach. Multilevel models are typically evaluated using a "build-up" approach (De Boeck & Wilson, 2004; Raudenbush & Bryk, 2002). An "empty" model (with no person-level or item-level predictors, only intercepts) is evaluated first. Predictors are added only in subsequent models, after the baseline model has been evaluated.

In a baseline model of the SCM, one could imagine that every item may have a unique level of difficulty (likelihood of being correctly answered) and every person may have a unique ability (likelihood of correctly answering). This is the concept of a "random intercept." The baseline EIRT model would allow for items to have different difficulties (random intercepts at item-level) and persons to have different abilities (random intercepts at person-level). Alternatively, if item difficulties and person abilities are the same across all items and persons, then the random intercepts reduce to standard regression intercepts. Similarly, random intercepts for solution times would be included in such a baseline model, and because models are evaluated with bivariate outcomes, covariances between solution time and accuracy intercepts would be evaluated.

A subsequent model would allow for the estimation of the effect of strategy usage on solution time and accuracy. The effect of strategy usage also might differ for different items and different persons. This is the concept of a random slope: the effect of strategy might be higher or lower than the average, both for persons and for items. Min-counting, for example, on some items might be more costly than it would be on other items (i.e., taking longer or decreasing accuracy), and likewise, min-counting for some persons might be more costly than it would be for other persons (i.e., increasing solution time or decreasing the likelihood of producing a correct answer). Alternatively, if the effect of strategy is constant, then the random slope effect reduces to a standard regression.

Accordingly, models were built in the following sequence: Model 1 was the baseline and was concerned with significant variance in random intercepts (i.e., the extent to which variance in outcomes, solution time and accuracy, could be explained by both item and person levels). Specifically, the baseline model examined the hypotheses that (a) items varied significantly from each other, (b) persons varied significantly from each other, and (c) in both solution times and accuracies. Given that accuracy demonstrated ceiling effects for this population, the extent to which items and persons might significantly vary from one another in accuracy was of particular concern in evaluating the baseline model.

Model 2 built upon Model 1 with the addition of random slopes for the use of the min-counting strategy; other strategies could have been modeled here, and min-counting was selected because it was frequently used across all addition problem types. Furthermore, because min-counting was the least sophisticated among the frequently used strategies, we anticipated that it might be useful for identifying adolescents who were struggling with mathematics. Model 2 estimated the extent to which the use of min-counting differed across both items and persons. This model included variance in random slopes, that is the extent to which of min-counting (vs. all other strategies) predicted solution time and accuracy across both items and persons. In subsequent analyses, item and person factor scores estimated from Model 2 were used to assess convergent validity with measures of broad math achievement, and discriminant validity with measures of reading ability. Table 4 presents a summary of results for all models, and Figures 1 and 2 display model schematics. Model equations are presented in Appendix A.

Model 1: Random intercepts only. The baseline model indicated significant variance in solution time and accuracy at both item and person levels.

Significant variance in solution time. Across items and persons, solution times varied by about 3 s on average (residual variance, σ^2_{ip} = 10. s^2, σ_{ip} = 3.19 s, $p < .001$). Items varied significantly in their solution times (item solution time intercept standard deviation, τ_{0i} = 2.79 s, $p < .001$). People varied significantly in their solution times (person solution time intercept standard deviation, τ_{0p} = 1.80 s, $p < .001$), but the average person took approximately 4 s to solve the average item (person solution time grand mean, γ_{00} = 4.48, $p < .001$). Thus, results from the baseline model indicated that both item- and person-level variance contributed to variance in solution times, justifying use of random intercepts. Collapsing across items (e.g., interpreting an average solution time mean for "simple addition problems") or collapsing across people (e.g., interpreting only a mean score for "adolescents") would be ignoring important sources of variance. This variance is commonly quantified by calculating an intraclass correlation (ICC) for a baseline, random intercept only model (Raudenbush & Bryk, 2002). ICCs were calculated as follows (note that all model parameters, represented here with Greek characters, are defined in Appendix A):

$$ICC_{items} = \frac{\tau^2_{0i}}{\tau^2_{0i} + \tau^2_{0p} + \sigma^2_{ip}} = \frac{7.794}{7.794 + 3.223 + 10.169} = .3679$$

Thus, approximately 37% of the variance in solution times was accounted for by differences between items, and

$$ICC_{persons} = \frac{\tau^2_{0p}}{\tau^2_{0i} + \tau^2_{0p} + \sigma^2_{ip}} = \frac{3.223}{7.794 + 3.223 + 10.169} = .1521$$

approximately 15% of the variance in solution times was accounted for by differences between people.

Significant variance in accuracy. Mplus 7 does not allow for the estimation of categorical outcomes at the within-level of cross-

Table 4
All Models Tested: Model Coefficient Parameters for Random and Fixed Effects

	Model 1		Model 2	
Effects	Solution time	Accuracy	Solution time	Accuracy
Person random effects				
Intercept/threshold	4.48 (.46)	−1.82 (.11)	4.15 (.37)	−1.90 (.08)
Intercept variance	3.22 (.39)	.24 (.05)	5.63 (.68)	.22 (.05)
Slope (the effect of counting)			2.59 (.57)	− .25 (.15)
Slope variance			3.61 (.59)	.28 (.15)
Item random effects				
Intercept variance	7.79 (2.58)	.23 (.08)	5.86 (1.67)	.25 (.09)
Slope variance			13.61 (4.18)	.19 (.10)
Error variance	10.17 (.21)	1.00 (fixed)	8.84 (.18)	1.00 (fixed)
Model fit				
Free parameters	7		25	

Note. The scales of solution time and accuracy are based on solution time and model probit metrics (similar to *z*-scores). Higher solution times mean slower problem-solving. Higher accuracy values mean higher likelihoods of a correct answer (i.e., easier items or more skilled problem solvers). Model covariance parameters of interest are presented and discussed in text.

classified models (in this study, the within-level is solution time and accuracy responses, fully cross-classified in items and persons); thus, estimating residuals for response accuracy was not possible (within-level residual variance is constrained equal to 1, as is conventional in probit regression). However, estimation of variance in accuracy was possible at item and person levels. Items varied significantly in their predicted accuracy (item accuracy intercept variance, $\tau_{0i}^2 = .23$, $p < .001$). People varied significantly in their predicted accuracy (person accuracy intercept variance, $\tau_{0p}^2 = .24$, $p < .001$), but the average person had a .97 probability of correctly solving the average item (person accuracy threshold, $\gamma_{00} = -1.82$, $p < .001$). Thus, results from the baseline model indicated that both item- and person-level variance contributed to variance in accuracy, justifying the use of random intercepts. Collapsing across items or across people would be ignoring important sources of variance. Approximately 16% of the variance in accuracy was accounted for by differences between items, and approximately 16% of the variance in accuracy was accounted for by differences between people.

Model 1 was re-examined for replicability using the second sample of random twin partners (singletons). Results for the second sample were largely consistent with those from the first. Results from replicability model testing are available in Appendix B.

Model 2: Random intercepts and random slopes for min-counting. The model indicated that both random intercepts and random slopes for min-counting had significant variance in solution time and accuracy at both item and person levels.

Significant variance in solution time. Across items and people, solution times varied by about 3 s on average (residual variance, $\sigma_{ip}^2 = 8.84$ s^2, $\sigma_{ip} = 2.97$ s, $p < .001$). Items varied significantly in their solution times (item solution time intercept standard deviation, $\tau_{0i} = 2.42$ s, $p < .001$). The use of min-counting varied in its effect on item solution times (item solution time slope standard deviation, $\tau_{1i} = 3.70$ s, $p < .001$). Note that solution times and the effects of min-counting on solution times are linearly related. For items, this relation is described by the covariance term $\tau_{10i} = 7.34$, or, in the more familiar $r = .81$. In other words, items that took longer to solve overall tended to take even longer to solve if min-counting strategy was used. These results indicated that both a random intercept and a random slope for min-counting captured significant variance (and covariance) at item-level.

People also varied significantly in their solution times (person solution time intercept standard deviation, $\tau_{0p} = 2.37$ s, $p < .001$), but on average, across all strategies used, people took approximately 4 s to solve each problem (person solution time grand mean, $\gamma_{00} = 4.15$, $p < .001$). The overall effect of min-counting (relative to the use of noncounting strategies such as retrieval) was about a 3 s increase in solution times (person solution time slope mean, $\gamma_{10} = 2.59$, $p < .001$). However, the use of min-counting strategies varied significantly across people (person solution time slope standard deviation, $\tau_{11} = 1.90$ s, $p < .001$), meaning there was significant variation in adolescents' efficiency of using min-counting. Interestingly, the significant covariance between the random intercept and slope ($\tau_{10p} = -4.02$; in the more familiar standardized form, $r = -.89$, $p < .001$) indicates that people who were slower overall were relatively fast when using the min strategy, whereas people who were faster overall tended to be slower when using min-counting. Taken together, the significant variance in random slopes for min-counting indicated that use of this strategy accounted for significant variance in solution times across items and persons.

Significant variance in accuracy. Both a random intercept and a random slope for min-counting captured significant variance in accuracy at the item-level. Items varied significantly in their predicted accuracy (i.e., their difficulty; item accuracy intercept variance, $\tau_{0i}^2 = .25$, $p < .001$). Items also differed in the probability that use of min-counting would yield the correct answer (item accuracy slope variance, $\tau_{1i}^2 = .19$, $p < .001$). However, item difficulty and the probability of generating the correct answer using min-counting were not significantly related ($\tau_{10i} = .05$, or, in the more familiar $r = .22$, $p = .22$). In other words, although the likelihood of generating a correct response using min-counting varied across items, this was unrelated to item difficulty.

People also varied significantly in their accuracy (person accuracy intercept variance, $\tau_{0p}^2 = .22$, $p < .001$), but the average person had a .97 probability of correctly solving the average item (person accuracy threshold, $\gamma_{00} = -1.90$, $p < .001$). Overall, for

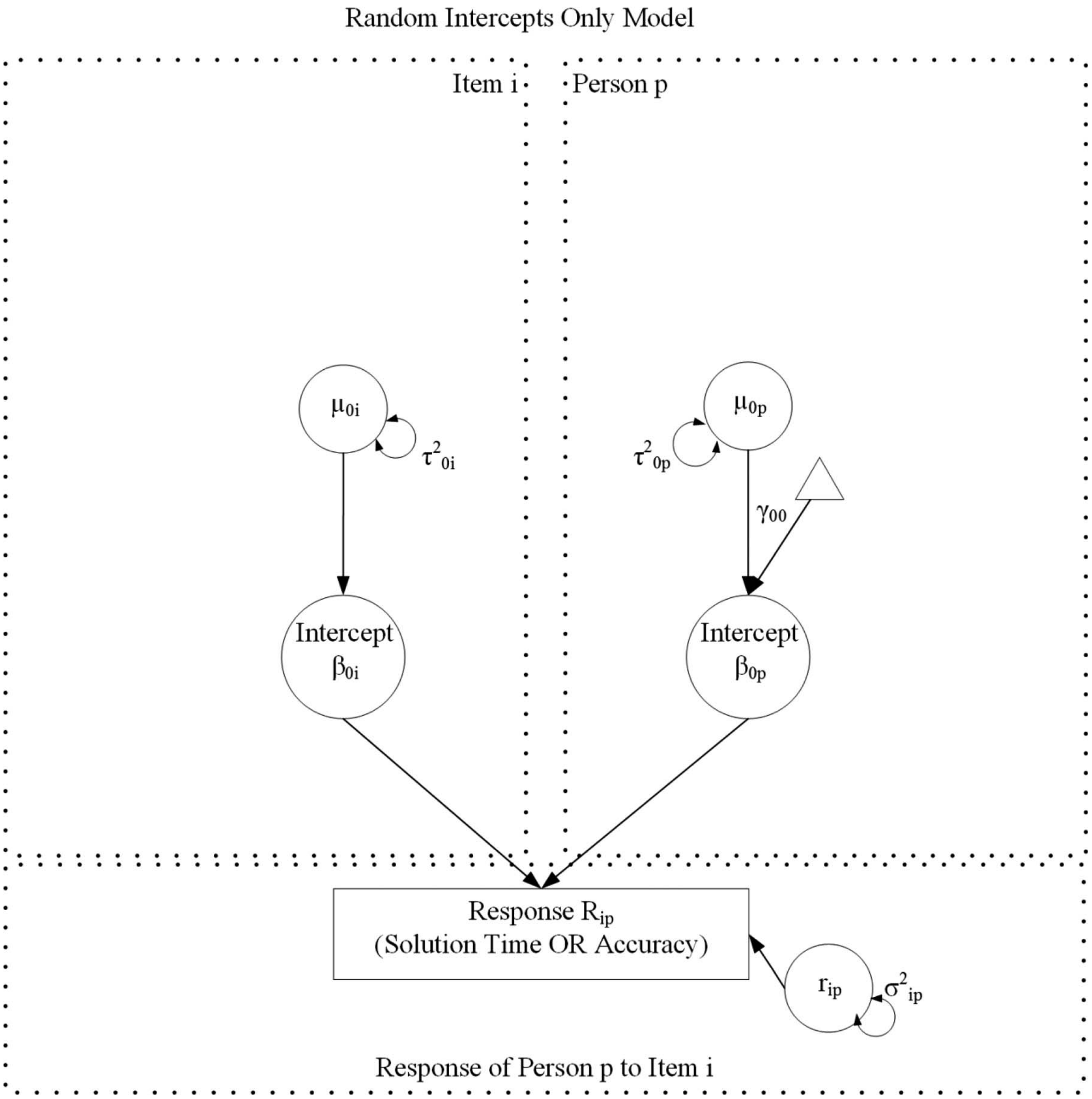


Figure 1. Explanatory item response theory Model 1 schematic. See the Technical Appendices for a full explanation of model symbols and equations. Note that this simple model schematic does not display the full, bivariate model under consideration (i.e., the schematic demonstrates the cross-classified mixed effects nature of the model in the current study; however, both solution time and accuracy responses are not displayed along with associated model covariance parameters).

the average person, the use of min-counting did not significantly impact the likelihood of a correct answer relative to use of other strategies (person accuracy slope mean, $\gamma_{10} = -.25$, $p = .07$). However, there were still significant individual differences in accuracy when using min-counting (person accuracy slope variance, $\tau^2_{1p} = .28$, $p < .001$). A nonsignificant covariance between the random intercept and slope ($\tau_{10p} = -.01$; in the more familiar $r = -.03$, $p = .43$) indicated that people who were less accurate overall were just as inaccurate when using min-counting; that is, they were just as likely to commit an error using min-counting as they were to commit an error using any other strategy.

Model 2 was reexamined for replicability using the second sample of random singletons. Again, the results were largely consistent with these results (see Appendix B). Similarly, an additional model including age as a fixed person effect was evaluated (i.e., controlling for age). Not surprisingly, older adolescents solved problems slightly faster than younger adolescents and tended to be slightly less efficient if using min-counting. All other model parameters demonstrated the same trends at similar magnitudes. Furthermore, despite the introduction of four additional free parameters for age effects, the model residual variances did not appreciably reduce, suggesting that this additional model complexity did not improve model estimates (95% CI [8.50, 9.20] and 95% CI [8.50, 9.17] for Model 2 and Model 2 with age, respectively). Thus, because the specification of item and person-level effects was beyond the

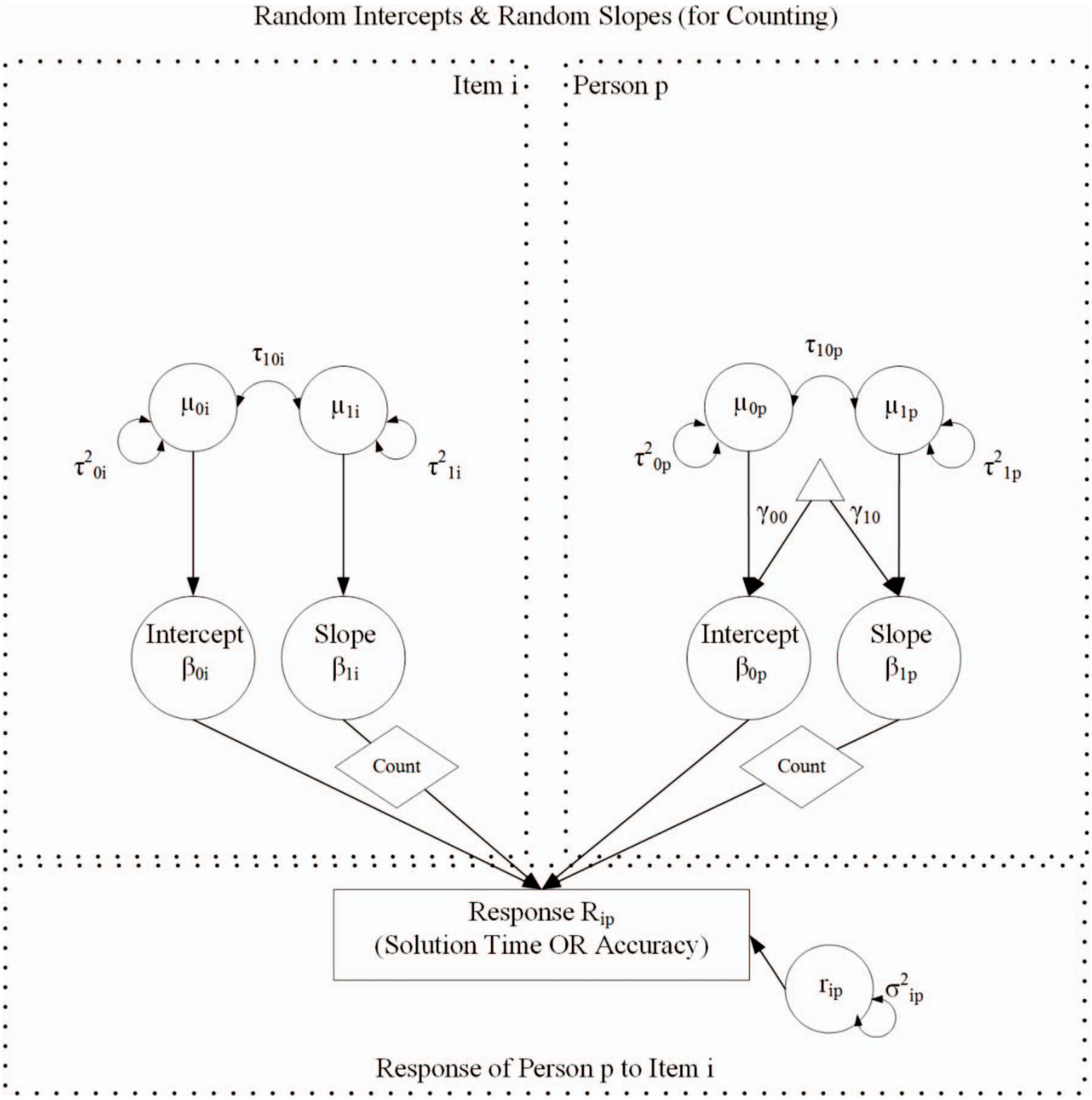


Figure 2. Explanatory item response theory Model 2 schematic. See the Technical Appendices for a full explanation of model symbols and equations. Note that this simple model schematic does not display the full, bivariate model under consideration (i.e., the schematic demonstrates the cross-classified mixed effects nature of the model in the current study; however, both solution time and accuracy responses are not displayed along with associated model covariance parameters).

scope of the current study, only the general Model 2 (across the adolescent stage of development) is presented here.

Examining EIRT Model Parameters for Individual Differences in Min-Counting

The average person could solve the average item in approximately 4.15 s with a probability of success = .97. These estimates are the EIRT intercept fixed effects, and they represent the typical adolescent performance on the addition problems across all strategies. However, both item and person intercepts had significant variance around these averages (i.e., significant random variance).

Similarly, the typical adolescent using a min-counting strategy took an average of 2.59 s longer to solve the average item, and on average, the use of min-counting did not compromise accuracy (probability of correct response = .95). These estimates are the EIRT slope fixed effects for the use of min-counting. However, again, there was significant variance in these slopes for both items and persons (i.e., significant random slope effects).

To demonstrate the utility of the EIRT modeling approach for capturing individual differences in problem-solving, factor scores (representing deviations from these averages) were considered with respect to both items and persons. Individual differences in item and person intercepts and slopes are considered, and extreme cases (representing individual persons and items for which model predicted effects were very different from peers) are highlighted in the sections that follow.

Individual differences in item intercepts. Scatter plots of item intercepts for the 30 addition problems are shown in Figure 3A (solution time) and 3B (accuracy). To illustrate, across persons, Items 22 (17 + 75) and 29 (65 + 28) were the most difficult. Item 22 had the second highest average solution time of 9.61 s and had the lowest average likelihood of correct response (probability = .82). Item 29 had the highest average solution time of 10.28 s and had the second lowest probability of a correct response (probability = .87). Conversely, Items 1 (3 + 6) and 2 (5 + 3) were the easiest items to solve (with averages of 2.23 s to solve and probability of success = .99, and 2.18 s and probability of success = .99, respectively).

Individual differences in item slopes. Figure 3C and 3D display scatter plots of item slopes for solution time and accuracy, respectively. In general, using min-counting was particularly maladaptive as problem sizes increased. For example, min-counting was a particularly poor choice for Item 29 (65 + 28), one of the most difficult problems on the assessment. Use of min-counting to solve this problem increased solution times by an average of 14.61 s with a probability of success = .56.

Individual differences in person intercepts. Figure 4A and 4B display intercepts for person-level solution time and accuracy, respectively. In general, most people were within two standard deviations of the average solution time and the average probability of success. However, there were several adolescents who took significantly longer than average to solve problems (deviated from average solution time by more than 2.5 *SD*s), and there were several adolescents who were significantly less likely to solve problems correctly (deviated from average accuracy by more than 2.5 *SD*s). For example, Participants 81 and 412 took an average of 22.85 s and 19.84 s, respectively, to solve the average problem. Participants 409 and 717 had an average probability of success, $p = .74$ and $p = .75$, respectively. Given the significant covariance between solution time and accuracy, it was not surprising that many of the slower adolescents were also the less accurate ones. For example, Participant 412 had a significantly lower than average probability of success, $p = .80$. Though there was significant variance among the typically performing participants, adolescents with these more extreme performances may have represented a

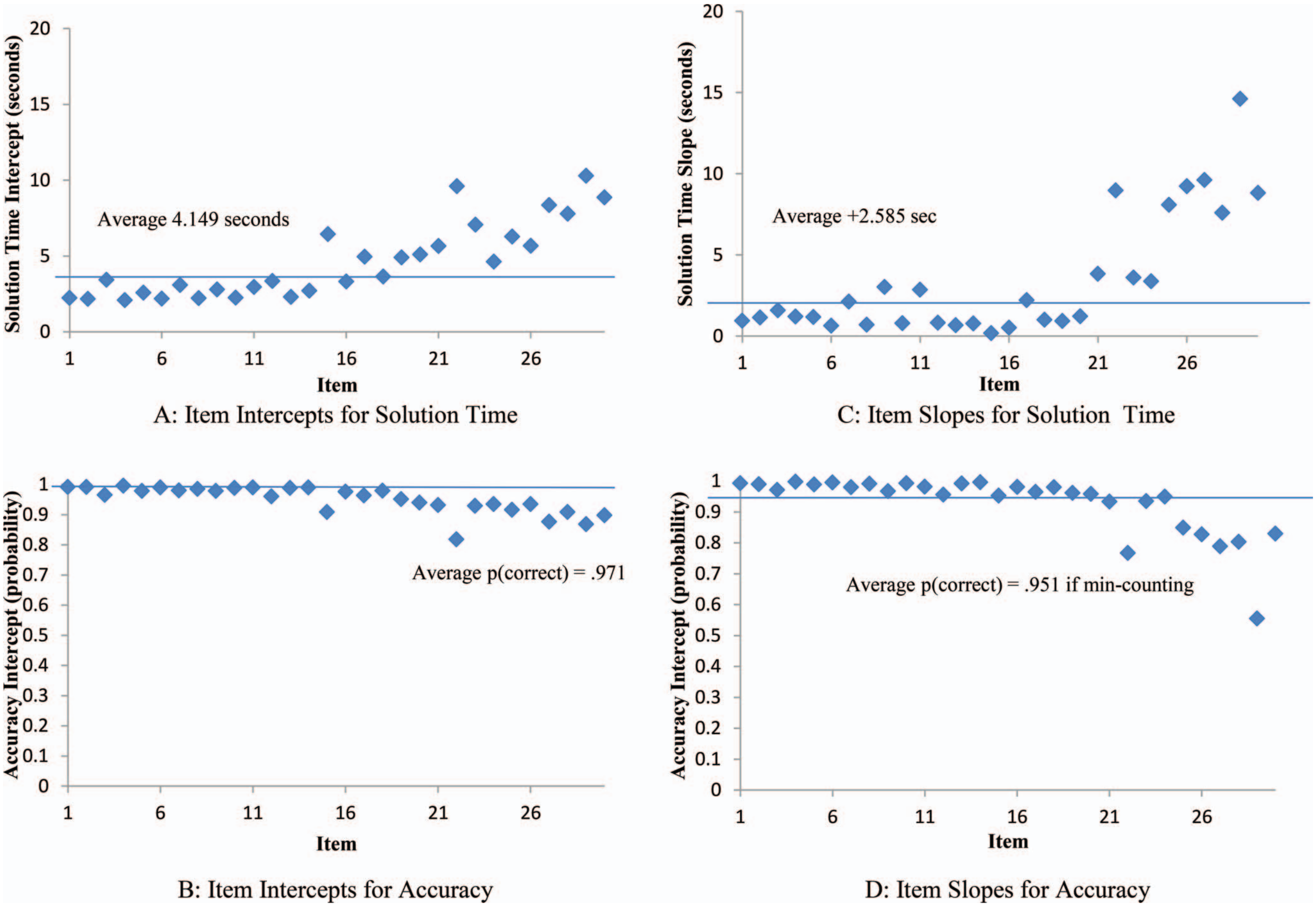
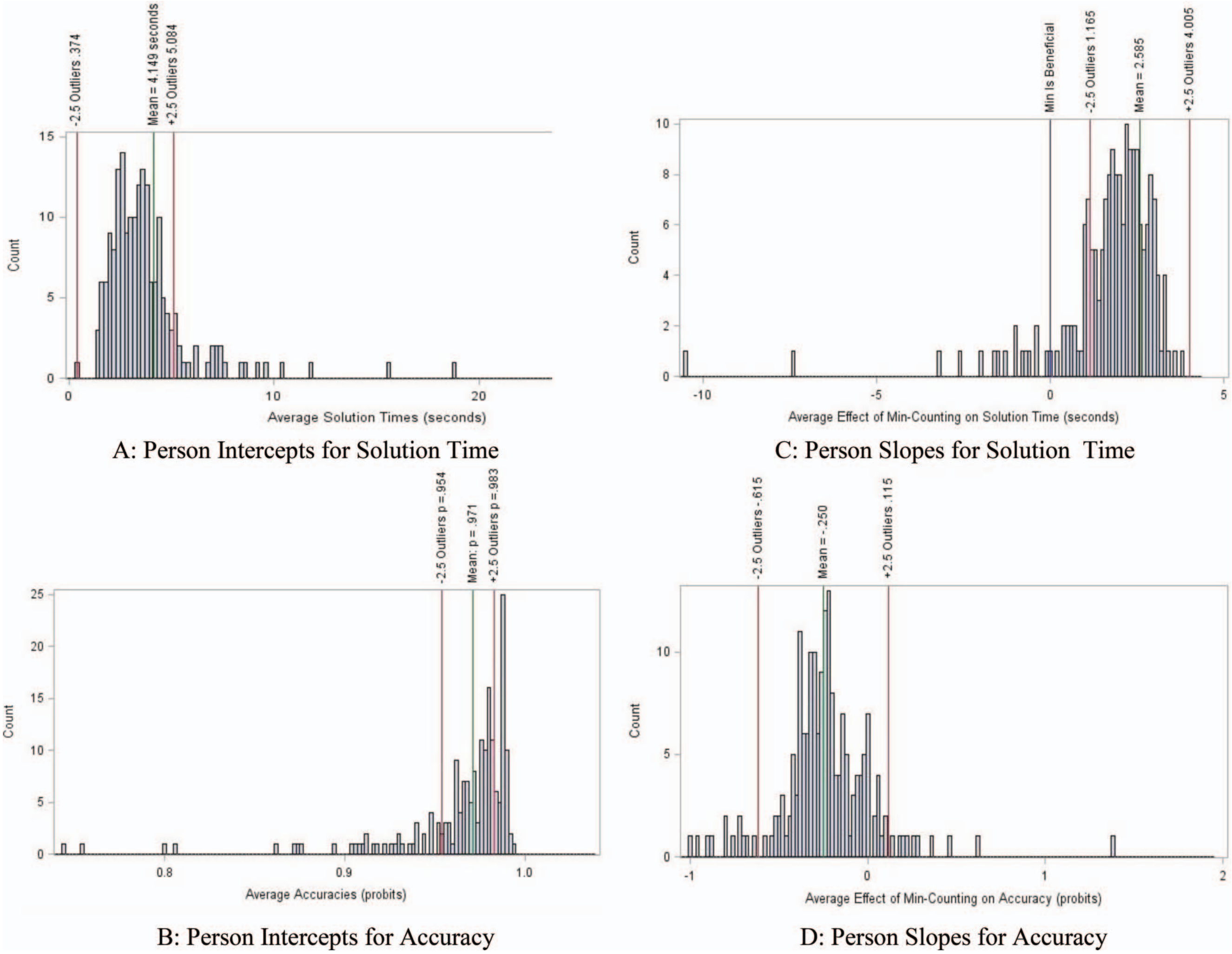


Figure 3. Item intercepts and slopes for solution time and accuracy. Items 1 to 14, inclusive are problems with single-digit addends (e.g., 5 + 3); Items 15 to 20, inclusive are problems with one single-digit and one double-digit (e.g., 15 + 7); problems 21 to 30, inclusive, are problems with two double-digit addends (e.g., 23 + 54). See the online article for the color version of this figure.



A: Person Intercepts for Solution Time

C: Person Slopes for Solution Time

B: Person Intercepts for Accuracy

D: Person Slopes for Accuracy

Figure 4. Person intercepts and slopes for solution time and accuracy. See the online article for the color version of this figure.

different population of problem-solver. The model predicted effects for using min-counting had significant implications for these adolescents and are considered next.

Individual differences in person slopes. Although most adolescents were not predicted to benefit from using min-counting in place of other strategies, some did. Figures 4C and 4D display individual differences for person-level min-counting slopes for solution time and accuracy, respectively. Figure 5 displays the bivariate distributions of min-counting effects. For most participants, use of min-counting was associated with longer solution times and more errors. For example, Participant 148 was slowed by min-counting an average of 3.82 s, and probability of success was reduced from $p = .994$ to $p = .989$. Most participants followed this pattern and used min-counting rather infrequently.

However, for some adolescents, the use of min-counting was predicted to be extremely beneficial for solution time alone ($n = 9$), and for other adolescents min-counting was predicted to be extremely beneficial for both solution time and accuracy ($n = 8$). Thus, these 17 adolescents were model predicted to experience extreme min-counting benefits, deviating substantially from their peers. For example, both of the slowest participants (Participants 81 and 412 mentioned above) were predicted to benefit from min-counting, reducing their average solution times by 10.47 s and 7.38 s, respectively. For both of these participants, min-counting was also predicted to significantly increase accuracy (e.g., from .91 to .99 for Participant 81).

Differences between model predicted "min-benefiters" and their peers. Surprisingly, those adolescents who were model predicted to benefit the most from min-counting used this strategy significantly less often than their peers ($t(186) = -4.51$, $p < .001$), suggesting maladaptive strategy choices.[1] The benefits of min-counting emerged because it is developmentally advanced compared to the counting strategies otherwise used by these students, for example, sum counting, finger counting, max-counting; $t(186) = 5.80$, $p < .001$. Put another way, those adolescents for

[1] Because multiple pairwise comparisons were considered, all reported *t*-test results represent the Tukey-Kramer adjusted values.

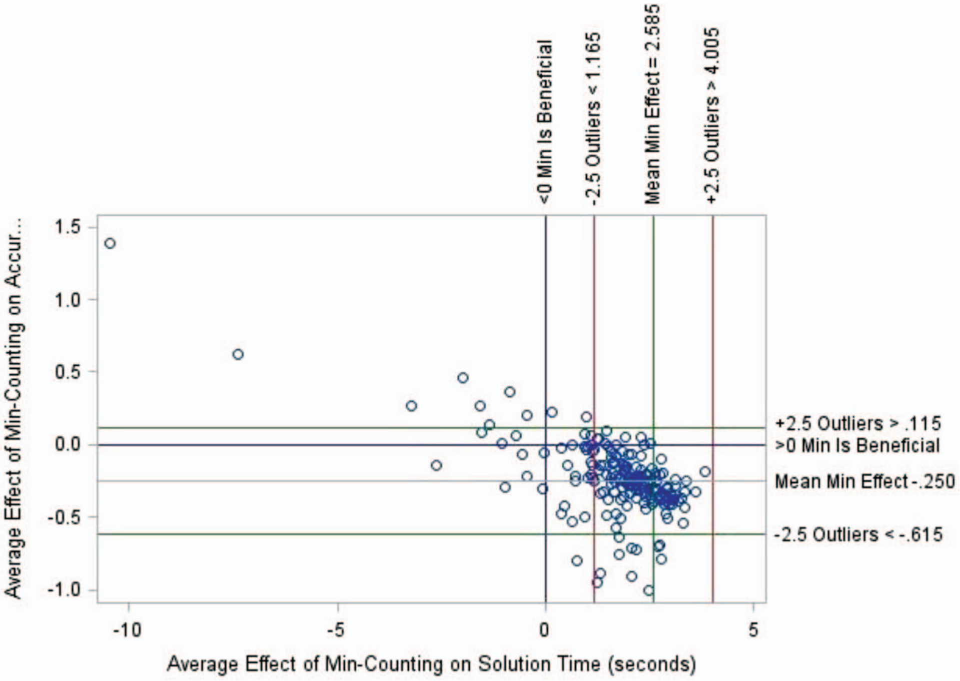


Figure 5. Accuracy BY Solution Time Slope Bivariate Distribution. See the online article for the color version of this figure.

whom the EIRT model predicted extreme beneficial effects of min-counting tended to rely on the especially immature (given their age) sum and max counting strategies and even finger counting to solve addition problems, whereas their peers rarely used these strategies.

Importantly, the model-predicted "min-benefiters" scored significantly lower than their peers in their broader mathematics achievement. These adolescents were about one standard deviation below average standard math achievement scores; calculations (mean standard score = 86.70), *t*(185) = 15.78, *p* < .001; applied problem-solving (mean standard score = 89.50), *t*(181) = 16.87, *p* < .001; math fluency (mean standard score = 85.12), *t*(186) = 18.93, *p* < .001; and broad mathematics (mean standard score = 83.75), *t*(180) = 21.74, *p* < .001.

Those who benefited in terms of solution time only or both solution time and accuracy were significantly different from their peers in frequency of min-counting, *t*(185) = −2.83, *p* = .01 and *t*(185) = −3.72, *p* < .001, respectively, but they were not significantly different from each other, *t*(185) = −.78, *p* = .72. However, those for whom the model predicted the most extreme benefits (in both solution time and accuracy) used naïve counting strategies significantly more than the solution time only benefiters, *t*(185) = 3.19, *p* = .01. Though these "dual benefiters" were more developmentally naïve than their "solution time only benefiter" counterparts, the two groups were not significantly different from each other in their proficiencies with calculations, *t*(184) = −1.97, *p* = .97, applied problems, *t*(180) = 4.00, *p* = .77, math fluency, *t*(185) = 1.40, *p* = .99, or broad mathematics, *t*(179) = .75, *p* = .99.

Examining EIRT Model Parameters for Convergent and Discriminant Validity

Though it was possible to examine group differences for min-benefiters, as Figures 4C, 4D, and 5 indicate, the distributions of EIRT model estimates were in fact continuous. Post hoc regression analyses tested the hypothesis that addition strategy modeling parameters (treated as continuous predictors) were predictive of broad math ability but less so of reading efficiency (see Table 5). Strategy predictors were examined using simultaneous entry. None of the models tested demonstrated issues of (non)linearity, and examination of residuals did not reveal any issues with non-normality, heteroscedasticity, or autocorrelation. However, given the strong correlations between the EIRT model intercepts and slopes, which were also evident in the replication sample, multicollinearity between predictors was a concern. Unsurprisingly, larger than desirable variance inflation factors (VIF > 10) and smaller than desirable tolerance values (tolerance < .20) were consistently noted for the solution time intercept and the accuracy intercept across models tested (O'Brien, 2007). Therefore, post hoc regression analyses did not include these addition strategy modeling parameters as predictors of broad math and reading outcomes. (Note that the "Intercept" values in Table 5 reflect intercepts for corresponding post hoc regression models, e.g., the average standard score of Broad Math, all other things being equal, and they do not reflect solution time or accuracy intercept values from the central EIRT model of the current study.)

The EIRT model parameters evidenced convergent validity with the Woodcock-Johnson Broad Math ability subscale and were predictive of performance on more narrow measures of mathemat-

Table 5
Convergent and Discriminant Validity Regression Analyses

Variable	Broad math ability				Calculations				Applied problems				Math fluency				Reading efficiency			
	B	*SE*	*t*-value	*p*	*B*	*SE*	*t*-value	*p*	*B*	*SE*	*t*-value	*p*	*B*	*SE*	*t*-value	*p*	*B*	*SE*	*t*-value	*p*
Intercept	45.73	13.37	3.42	<.001	55.84	14.65	3.81	<.001	52.58	9.83	5.35	<.001	49.92	16.38	3.05	.003	76.52	13.33	5.74	<.001
Soln. Tm. Slope	2.74	.66	4.12	<.001	2.29	.78	2.95	.004	2.06	.49	4.21	<.001	2.92	.87	3.36	.001	−.76	.67	−1.13	.26
Acc. Slope	61.96	13.94	4.45	<.001	48.86	15.25	3.20	.002	56.06	10.25	5.47	<.001	54.37	17.05	3.19	.002	26.28	13.87	1.89	.06
Naive count ratio	−23.00	4.76	−4.84	<.001	−20.10	5.61	−3.58	<.001	−18.27	3.49	−5.23	<.001	−15.36	6.27	−2.45	.02	−12.34	4.56	−2.70	.01
Model fit																				
F value	$F(3, 176) = 42.88, p < .001$				$F(3, 181) = 21.93, p < .001$				$F(3, 177) = 51.65, p < .001$				$F(3, 182) = 18.20, p < .001$				$F(3, 137) = 4.22, p = .01$			
R^2	.42				.27				.47				.23				.08			

Note. Soln. = solution; Tm. = time; Acc. = accuracy.

ical competence (see Table 5). Addition strategy solution time slope was a significant, positive predictor of broad math ability, $B = 2.74$, $t(176) = 4.12$, $p < .001$. Thus, for every second that min-counting increased solution time there was an average 2.74 standard score increase in broad math scores. Addition strategy accuracy slope was also a significant, positive predictor of broad math ability, $B = 61.96$, $t(176) = 4.45$, $p < .001$. For every .01 unit increase in probability associated with min-counting, there was an average .62 standard unit increase in broad math scores. Finally, the ratio of addition strategy items upon which immature counting strategies (i.e., nonmin-counting strategies such as finger counting all) were used was a significant, negative predictor of broad math ability, $B = -23.00$, $t(176) = -4.84$, $p < .001$. For every .01 unit increase in the proportion of items on which naïve counting was used, there was an average .23 unit decrease in broad math standard scores. The strategy variables accounted for a significant portion of variance in broad math ability standard scores, $R^2 = .42$, $F(3, 176) = 42.88$, $p < .001$.

The EIRT model parameters also evidenced discriminant validity with reading efficiency. Neither the solution time slope nor the accuracy slope was a significant predictor of reading efficiency $B = -.76$, $t(137) = -1.13$, $p = .26$, and $B = 26.28$, $t(137) = 1.89$, $p = .06$, respectively. Interestingly, the proportion of items on which immature counting strategies were used was a significant predictor of reading efficiency, $B = -12.34$, $t(137) = -2.70$, $p = .01$. For every .01 unit increase in the proportion of items on which immature counting strategies were used, there was an average .12 unit decrease in reading efficiency scores. Still, as was expected, this model explained very little variance in reading efficiency $R^2 = .08$, $F(3, 137) = 4.22$, $p = .01$.

Summary of Major Findings

Results from Model 1 indicated that differences between both items and persons contributed to significant variances in solution time and accuracy outcomes. Intraclass correlations indicated that failing to model these differences (random intercepts) would be ignoring between 12 and 37% of the variances in adolescent problem-solvers' speed and accuracy in solving addition problems.

Model 2 showed that use of the min-counting strategy affected solution times and accuracies but did so differently for different items and persons. Thus, as the SCM postulates, the costs and benefits of using min-counting relative to other strategies varied depending on the problem-solver and item being solved. The magnitudes and directions of these results were robust and replicated across random samples of singleton participants.

Model 2 parameters also revealed significant individual differences in both items and persons for solution times and accuracies, and that some adolescents deviated substantively from their peers by showing atypical strategy choices. These were adolescents for whom min-counting was an efficient strategy, given they often used sum counting and other immature strategies, but they did not use it consistently. More broadly, the EIRT parameters demonstrated convergent and discriminant validity; they were predictive of broad mathematics but not reading achievement. Taken together, these results indicate that the EIRT model parameters for a simple addition strategy task provided estimates for individual adolescents that were reliable, valid, and useful for identifying students with broad difficulties with mathematics.

Discussion

The current study examined the SCM (Siegler & Robinson, 1982; Siegler & Shrager, 1984) from an individual difference perspective using EIRT modeling techniques; specifically, demonstrating both item-level and person-level variation in the strategy choices and solution times of adolescent problem-solvers. Not only did persons and items differ from each other, but the effect of the min-counting strategy had different levels of impact for different persons on different items. In addition, EIRT model estimates of individual differences were examined for their utility in identifying adolescents with unusual patterns of strategy choices and solution times, relative to adolescents generally, and for their convergent and discriminant validity with measures of broad academic achievement.

Results supported a central tenet of the SCM, that variance in both items and persons can affect problem-solving outcomes. Indeed, significant variance in both item and person-level intercepts indicated that collapsing across either of these levels could yield highly overgeneralized results, confirming Siegler's (1987b) early caveats about averaging across items and extending this to averaging across people. The EIRT framework of explanatory measurement provided an ideal outlet for examining the multilevel variance predicted by the SCM.

The min-counting strategy was used to solve relatively simple addition problems with surprising frequency in this sample of adolescents, but the efficacy of its use depended upon both items and persons. Min-counting became increasingly maladaptive as problem sizes increased. Most participants were slower when using min-counting relative to the other strategies, such as retrieval or decomposition, but not significantly more or less likely to correctly answer problems. However, for some problem-solvers, min-counting provided significant benefits compared to the less sophisticated strategies (e.g., sum counting) they frequently employed.

The "min-benefiters" identified by this EIRT model would have been obscured using a more traditional approach to SCM evaluation because their strategy choice pattern would have been collapsed into the pattern of their typically developing peers. In other words, the overall (across the entire sample) use of unsophisticated counting strategies (e.g., sum counting) would have seemed infrequent and the benefits of min-counting for some adolescents would not have become apparent. These adolescents were not members of an atypical population, and none had been identified with MLD. Indeed, the EIRT model parameters were distributed continuously, and there were no apparent clusters of persons at the tail end of the distribution, as is generally indicative of distinct populations. Rather, these "min-benefiters" appeared to be adolescents who are sometimes called "garden variety low achievers." They were consistently about one standard deviation below average in their broad mathematics achievement, not low enough to be labeled as MLD in most settings but not high enough to meet grade level standards. In any case, their use of unsophisticated counting strategies and the attendant benefits of min-counting are consistent with studies of low achieving elementary schoolchildren; that is, they use the same types of strategies in problem solving as typically achieving children, but the strategy mix is similar to that found in younger children (Geary, 1993). Taken together, these results indicate that a simple addition assessment may provide researchers and practitioners with valuable information about mathematical skill development, when the assessment is interpreted with a sound combination of theory (the SCM) and analytical approach (EIRT).

Implications for Individual Differences Research on Strategy Choice

Using the traditional approach to examine the SCM, one would assume that all participants were from a similar population of problem-solver and collapse data across items and/or persons to describe average effects (Siegler, 1987b). Such an approach has been useful for detailing group (e.g., comparing groups of low- and typically achieving children) or grade-level differences in the pattern of strategy choices (e.g., overall frequency of min-counting across problems) but does not fully capture item-level and person-level variation in these choices (see e.g., Bailey et al., 2012; Geary et al., 2004; Siegler, 1991). In other words, a traditional individual differences analysis based on the SCM would have used clustering analyses or criterion measures of population (e.g., math learning disability diagnosis) to collapse across individuals in "different" populations. In both of these analytic approaches, the natural variance between items and persons would have been largely obscured in favor of emphasizing universal trends in problem-solving for the identified group.

The EIRT approach used in the current study demonstrated, in keeping with predictions based on the SCM, that even typical problem-solvers have natural fluctuations in speed, accuracy, and strategy usage during problem solving. Some adolescents in the current study, however, demonstrated problem-solving trends that were very different from their peers. The more nuanced examination of individual differences in the current study led to the observation that counting (a) had not receded as a primary strategy for solving addition problems for many adolescents and (b) appeared to be adaptive for some adolescents but maladaptive for most. Integrating the behavioral universals and individual differences perspectives in complementary ways may have interesting implications for the SCM as it continues to evolve.

Implications for Identifying and Remediating Mathematical Difficulties

Most students in the United States do not achieve grade level proficiency in mathematics, and the percentages of students performing at grade level drop dramatically as they progress through school (Kelly et al., 2014; National Center for Education Statistics, 2013). The majority of these low achieving students are not identified with a specific mathematical learning disability. Their difficulties with mathematics are not necessarily extreme enough to warrant such a diagnosis, and simply falling below grade level standards may or may not be indicative of the difficulties with working memory, processing speed, fact retrieval, and procedural deficits associated with MLD (see, e.g., Compton, Fuchs, Fuchs, Lambert, & Hamlett, 2012; Geary, 1993). The "min-benefiters" identified in the current study were likely representative of this larger trend in mathematical (under)achievement.

However, the fact that the "min-benefiters" were typically developing adolescents who had not been identified with MLD does not mean they would not benefit from intervention. Indeed, the current results suggested they would likely benefit from instruction that fostered the more consistent use of min-counting in place of the less sophisticated strategies they often used (e.g., counting all fingers). The results from the current study suggested that not only would these adolescents benefit from mastering min-counting in proximal outcomes (solution times and accuracy on a simple addition assessment) but also in more distal outcomes (calculations, applied problems, math fluency, and broad math achievement). For students with garden variety mathematical difficulties, simply advancing to developmentally mature strategy selection follows from the SCM and may be an important intervention goal, one that is often overlooked in educational settings (but see Fuchs et al., 2013).

Limitations and Future Directions

Though the primary purpose of the current study was to examine the SCM using EIRT modeling techniques with a sample of adolescent problem-solvers, it should be noted that it is possible to build upon Model 2 by adding slopes for other types of strategies (e.g., retrieval, decomposition), item-level predictors (e.g., fixed effects for problem sum, item formatting), and/or person-level predictors (e.g., fixed effects for working memory, and/or gender). In particular, unpacking and disaggregating strategies to examine individual differences in use of decomposition and retrieval may provide valuable insights into problem-solving and could be particularly useful for identifying gifted problem-solvers. However, the current study was an initial attempt at evaluating the SCM using EIRT. Expanding the complexity of the current model to include more random effects for other strategies was beyond the scope of this study and likely underpowered given the current study's sample size. Future research should examine the effects of both immature strategies (e.g., sum-counting, min-counting) and more advanced strategies (e.g., retrieval, decomposition) in models with multiple slopes for strategy and sample sizes large enough to accommodate these effects.

Similarly, the sample in the current study represented typically developing adolescents (very few reported disabilities and these were generally mild disabilities such as attention problems), who were mostly White and largely from middle-class families. In addition, because the Bayes estimator can be sensitive to the presence of extreme cases, current model estimates may be inflated due to the presence of the "min-benefiters." Extreme cases were not excluded from the current study because they were (a) valid data points, and (b) of direct consequence to the individual differences research question. However, taken together, caution should be used in generalizing the specific results from the present study to other populations of problem-solvers. Future research should examine EIRT models with larger samples representing additional populations. In particular, given the wealth of research regarding the SCM's applicability to children with mathematical learning disabilities, researchers should extend this model to children with different learning profiles.

Conclusion

Both items and persons contributed significant variance to the addition problem-solving performance considered in this study, and the use of an EIRT framework allowed for the explicit testing of this central tenet of the SCM. The use of EIRT models also allowed for a more nuanced exploration of individual differences between items and persons, because each item and each person was allowed to vary from an average effect. Thus, the EIRT approach provided validations and unique insights into the individual differences in strategy choice during the period of adolescence. Of particular interest to researchers and practitioners are the findings that (a) even among typically developing adolescents, individual differences in strategy choice on a simple addition measure were meaningful indicators of ability, and (b) strategy choice provided important insights for remediating mathematics achievement difficulties.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59,* 390–412. http://dx.doi.org/10.1016/j.jml.2007.12.005

Bailey, D. H., Littlefield, A., & Geary, D. C. (2012). The codevelopment of skill at and preference for use of retrieval-based processes for solving addition problems: Individual and sex differences from first to sixth grades. *Journal of Experimental Child Psychology, 113,* 78–92. http://dx.doi.org/10.1016/j.jecp.2012.04.014

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105. http://dx.doi.org/10.1037/h0046016

Compton, D. L., Fuchs, L. S., Fuchs, D., Lambert, W., & Hamlett, C. (2012). The cognitive and academic profiles of reading and mathematics learning disabilities. *Journal of Learning Disabilities, 45,* 79–95. http://dx.doi.org/10.1177/0022219410393012

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer. http://dx.doi.org/10.1007/978-1-4757-3990-9

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197. http://dx.doi.org/10.1037/0033-2909.93.1.179

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8,* 430–457. http://dx.doi.org/10.1207/S15328007SEM0803_5

Fuchs, L. S., Geary, D. C., Compton, D. L., Fuchs, D., Schatschneider, C., Hamlett, C. L., . . . Changas, P. (2013). Effects of first-grade number knowledge tutoring with contrasting forms of practice. *Journal of Educational Psychology, 105,* 58–77. http://dx.doi.org/10.1037/a0030127

Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin, 114,* 345–362. http://dx.doi.org/10.1037/0033-2909.114.2.345

Geary, D. C., & Brown, S. C. (1991). Cognitive addition: Strategy choice and speed-of-processing differences in gifted, normal, and mathematically disabled children. *Developmental Psychology, 27,* 398–406. http://dx.doi.org/10.1037/0012-1649.27.3.398

Geary, D. C., Fan, L., & Bow-Thomas, C. C. (1992). Numerical cognition: Loci of ability differences comparing children from China and the United States. *Psychological Science, 3,* 180–185. http://dx.doi.org/10.1111/j.1467-9280.1992.tb00023.x

Geary, D. C., Frensch, P. A., & Wiley, J. G. (1993). Simple and complex mental subtraction: Strategy choice and speed-of-processing differences in younger and older adults. *Psychology and Aging, 8,* 242–256. http://dx.doi.org/10.1037/0882-7974.8.2.242

Geary, D. C., Hoard, M. K., & Bailey, D. H. (2012). Fact retrieval deficits in low achieving children and children with mathematical learning disability. *Journal of Learning Disabilities, 45,* 291–307. http://dx.doi.org/10.1177/0022219410392046

Geary, D. C., Hoard, M. K., Byrd-Craven, J., & DeSoto, M. C. (2004). Strategy choices in simple and complex addition: Contributions of working memory and counting knowledge for children with mathematical disability. *Journal of Experimental Child Psychology, 88,* 121–151. http://dx.doi.org/10.1016/j.jecp.2004.03.002

Geary, D. C., Hoard, M. K., Nugent, L., & Rouder, J. N. (2015). Individual differences in algebraic cognition: Relation to the approximate number and semantic memory systems. *Journal of Experimental Child Psychology, 140,* 211–227. http://dx.doi.org/10.1016/j.jecp.2015.07.010

Geary, D. C., Widaman, K. F., Little, T. D., & Cormier, P. (1987). Cognitive addition: Comparison of learning disabled and academically normal elementary school children. *Cognitive Development, 2,* 249–269. http://dx.doi.org/10.1016/S0885-2014(87)90075-X

Geary, D. C., & Wiley, J. G. (1991). Cognitive addition: Strategy choice and speed-of-processing differences in young and elderly adults. *Psychology and Aging, 6,* 474–483. http://dx.doi.org/10.1037/0882-7974.6.3.474

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap)—a metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics, 42,* 377–381.

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics, 23,* 117–128. http://dx.doi.org/10.3102/10769986023002117

Jordan, N. C., Hanich, L. B., & Kaplan, D. (2003). Arithmetic fact mastery in young children: A longitudinal investigation. *Journal of Experimental Child Psychology, 85,* 103–119. http://dx.doi.org/10.1016/S0022-0965(03)00032-8

Kelly, D., Xie, H., Nord, C. W., Jenkins, F., Chan, J. Y., & Kastberg, D. (2014). *Performance of U.S. 15-year-old students in mathematics, science, and reading literacy in an international context: First look at PISA 2012*. Retrieved from http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2014024

Kerkman, D. D., & Siegler, R. S. (1993). Individual differences and adaptive flexibility in lower-income children's strategy choices. *Learning and Individual Differences, 5,* 113–136. http://dx.doi.org/10.1016/1041-6080(93)90008-G

Kerkman, D. D., & Siegler, R. S. (1997). Measuring individual differences in children's addition strategy choices. *Learning and Individual Differences, 9,* 1–18. http://dx.doi.org/10.1016/S1041-6080(97)90017-0

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language, 94,* 305–315. http://dx.doi.org/10.1016/j.jml.2017.01.001

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock-Johnson III Tests of Achievement*. Itasca, IL: Riverside.

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Author.

National Center for Education Statistics. (2013). *The nation's report card: A first look: 2013 mathematics and reading* (NCES 2014–451). Washington, DC: Institute of Education Sciences, U.S. Department of Education. Retrieved from papers3://publication/uuid/45C8F3B4-8A8D-459C-A3C3-2D59E6429AA4

O'Brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity: International Journal of Methodology, 41,* 673–690. http://dx.doi.org/10.1007/s11135-006-9018-6

Petrill, S. A., Deater-Deckard, K., Thompson, L. A., Dethorne, L. S., & Schatschneider, C. (2006). Reading skills in early readers: Genetic and shared environmental influences. *Journal of Learning Disabilities, 39,* 48–55. http://dx.doi.org/10.1177/00222194060390010501

Qin, S., Cho, S., Chen, T., Rosenberg-Lee, M., Geary, D. C., & Menon, V. (2014). Hippocampal-neocortical functional reorganization underlies children's cognitive development. *Nature Neuroscience, 17,* 1263–1269. http://dx.doi.org/10.1038/nn.3788

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: SAGE Publications, Inc.

SAS Institute Inc. (2011). *Base SAS 9.3 procedures guide*. Cary, NC: Author.

Siegler, R. S. (1986). Unities across domains in children's strategy choices. In M. Perlmutter (Ed.), *Perspectives on intellectual development: The Minnesota symposia on child psychology* (Vol. 19, pp. 1–48). Hillsdale, NJ: Erlbaum Publishers.

Siegler, R. S. (1987a). Some general conclusions about children's strategy choice procedures. *International Journal of Psychology, 22,* 729–749. http://dx.doi.org/10.1080/00207598708246800

Siegler, R. S. (1987b). Strategy choices in subtraction. In J. Sloboda & D. Rogers (Eds.), *Cognitive processes in mathematics* (pp. 81–106). Oxford, UK: Oxford University Press.

Siegler, R. S. (1987c). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General, 116,* 250–264. http://dx.doi.org/10.1037/0096-3445.116.3.250

Siegler, R. S. (1988a). Individual differences in strategy choices: Good students, not-so-good students, and perfectionists. *Child Development, 59,* 833–851. http://dx.doi.org/10.2307/1130252

Siegler, R. S. (1988b). Strategy choice procedures and the development of multiplication skill. *Journal of Experimental Psychology: General, 117,* 258–275. http://dx.doi.org/10.1037/0096-3445.117.3.258

Siegler, R. S. (1991). Strategy choice and strategy discovery. *Learning and Instruction, 1,* 89–102. http://dx.doi.org/10.1016/0959-4752(91)90020-9

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York, NY: Oxford University Press.

Siegler, R. S., & Jenkins, E. (1989). *How children discover new strategies*. Hillsdale, NJ: Erlbaum Publishers.

Siegler, R. S., & McGilly, K. (1989). Strategy choices in children's time-telling. In I. Levin & D. Zakay (Eds.), *Advances in psychology: Time and human cognition: A lifespan perspective* (Vol. 59, pp. 185–218). New York, NY: Elsevier.

Siegler, R. S., & Robinson, M. (1982). The development of numerical understanding. *Advances in Child Development and Behavior, 16,* 241–312. http://dx.doi.org/10.1016/S0065-2407(08)60072-5

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills: The eighteenth annual carnegie symposium on cognition* (pp. 229–293). Hillsdale, NJ: Erlbaum Publishers.

Siegler, R. S., & Taraban, R. (1986). Conditions of applicability of a strategy choice model. *Cognitive Development, 1,* 31–51. http://dx.doi.org/10.1016/S0885-2014(86)80022-3

Supekar, K., Swigart, A. G., Tenison, C., Jolles, D. D., Rosenberg-Lee, M., Fuchs, L., & Menon, V. (2013). Neural predictors of individual differences in response to math tutoring in primary-grade school children. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 8230–8235. http://dx.doi.org/10.1073/pnas.1222154110

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of word reading efficiency*. Austin, TX: Pro-Ed. Publishing.

Widaman, K. F., Geary, D. C., Cormier, P., & Little, T. D. (1989). A componential model for mental addition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 898–919. http://dx.doi.org/10.1037/0278-7393.15.5.898

(*Appendices follow*)

Appendix A

Technical EIRT Model Equations

Model 1: Mixed Model Equation

$$R_{ip} = \beta_{0i} + \beta_{0p} + r_{ip}$$

where R_{ip} is *person p*'s response (solution time or accuracy) to *item I* and B_{0i} is the random intercept of *item i*, or *item i*'s average response across all people, allowed to vary across items such that response is item-specific.

$$\beta_{0i} = \mu_{0i};\ \mu_{0i} \sim N(0, \tau_{0i}^2)$$

μ_{0i} is the deviation of *item i* from the mean response of all items, distributed normally with $M = 0$ and variance $= \tau_{0i}^2$, thus for the average item, $B_{0i} = \mu_{0i} = 0$. B_{0p} is the random intercept of *person p*, or *person p*'s average response across all items, allowed to vary across people such that response is person-specific

$$\beta_{0p} = \gamma_{00} + \mu_{0p};\ \mu_{0p} \sim N(0, \tau_{0p}^2)$$

γ_{00} is the mean response of all people, the grand mean; μ_{0p} is the deviation of *person p* from the mean response of all people γ_{00}, and μ_{0p} is distributed normally with $M = 0$ and variance $= \tau_{0p}^2$, thus for the average person, $B_{0p} = \gamma_{00}$; and r_{ip} is the model residual for response, normally distributed with mean 0 and variance σ_{ip}^2.

$$r_{ip} \sim N(0, \tau_{ip}^2)$$

Thus, substituting the random intercept and slope equations in the mixed model equation:

$$R_{ip} = \mu_{0i} + \gamma_{00} + \mu_{0p} + r_{ip}$$

Model 2: Mixed Model Equation

$$R_{ip} = \beta_{0i} + \beta_{1i}C_{ip} + \beta_{0p} + \beta_{1p}C_{ip} + r_{ip}$$

Using the distributive property of multiplication, this equation can be rewritten as

$$R_{ip} = \beta_{0i} + \beta_{0p} + (\beta_{1i} + \beta_{1p})C_{ip} + r_{ip}$$

where R_{ip} is *person p*'s response (solution time or accuracy) to *item i* and B_{0i} is the random intercept of *item i*, or *item i*'s average response across all people, allowed to vary across items such that response is item-specific.

$$\beta_{0i} = \mu_{0i};\ \mu_{0i} \sim N(0, \tau_{0i}^2)$$

μ_{0i} is the deviation of *item i* from the mean response of all items, distributed normally with $M = 0$ and variance $= \tau_{0i}^2$, thus for the average item, $B_{0i} = \mu_{0i} = 0$. B_{1i} is the random slope of *item i*, the random effect of counting strategy usage on response, allowed to vary across items such that the effect of counting is item-specific.

$$\beta_{1i} = \mu_{1i};\ \mu_{1i} \sim N(0, \tau_{1i}^2)$$

μ_{1i} is the deviation of *item i* from the mean effect of counting on response across all items, distributed normally with $M = 0$ and variance $= \tau_{1i}^2$ and τ_{10i} is the covariance between the random intercept and the random slope, the extent to which the use of counting strategies tends to effect responses. B_{0p} is the random intercept of *person p*, or *person p*'s average response across all items, allowed to vary across people such that response is person-specific.

$$\beta_{0p} = \gamma_{00} + \mu_{0p};\ \mu_{0p} \sim N(0, \tau_{0p}^2)$$

γ_{00} is the mean response of all people, the grand mean, and μ_{0p} is the deviation of *person p* from the mean response of all people γ_{00}, and μ_{0p} is distributed normally with $M = 0$ and variance $= \tau_{0p}^2$, thus for the average person, $B_{0p} = \gamma_{00}$. B_{1p} is the random slope of *person p*, the random effect of counting strategy usage on *person p*'s response, allowed to vary across people such that the effect of counting is person-specific

$$\beta_{1p} = \gamma_{10} + \mu_{1p};\ \mu_{1p} \sim N(0, \tau_{1p}^2)$$

γ_{10} is the mean effect of counting strategy usage on response across all people, the grand mean; μ_{1p} is the deviation of *person p* from the mean effect of counting on response across all people (γ_{10}), distributed normally with $M = 0$ and variance $= \tau_{1p}^2$; and τ_{10p} is the covariance between the random intercept and the random slope, the extent to which the use of counting strategies tends to effect responses. C_{ip} is the term denoting whether *person p* counted on *item i* (0 if no counting strategy was used; 1 if a counting strategy was used) and r_{ip} is the model residual for response, normally distributed with mean 0 and variance σ_{ip}^2.

$$r_{ip} \sim N(0, \tau_{ip}^2)$$

Thus, substituting random intercept and slope equations in the mixed model . . .

$$R_{ip} = \mu_{0i} + \gamma_{00} + \mu_{0p} + (\mu_{1i} + \gamma_{10} + \mu_{1p})C_{ip} + r_{ip}$$

(*Appendices continue*)

Appendix B

Technical Model Replicability Testing

Model 1 Replicability

Across items and people, solution times varied by about 4 s on average (residual variance, $\sigma_{ip} = 3.59$ s, $p < .001$). Items varied significantly in their solution times (item solution time intercept standard deviation, $\tau_{0i} = 2.84$ s, $p < .001$). People varied significantly in their solution times (person solution time intercept standard deviation, $\tau_{0p} = 2.23$ s, $p < .001$), but the average person took approximately 4 s to solve the average item (person solution time grand mean, $\gamma_{00} = 3.93$, $p < .001$). Approximately 31% of the variance in solution times was accounted for by differences between items, and approximately 19% of the variance in solution times was accounted for by differences between people.

Items also varied significantly in their predicted accuracy (item accuracy intercept variance, $\tau^2_{0i} = .24$, $p < .001$). People varied significantly in their predicted accuracy (person accuracy intercept variance, $\tau^2_{0p} = .17$, $p < .001$), but the average person had a probability of .96 of correctly solving the average item (person accuracy threshold, $\gamma_{00} = -1.70$, $p < .001$). Approximately 17% of the variance in accuracy was accounted for by differences between items, and approximately 12% of the variance in accuracy was accounted for by differences between people. Taken together, these results also suggested that failing to model individual variance for both items and persons could lead to missing important sources of variance in both solution times and accuracies.

Model 2 Replicability

Across items and people, solution times varied by about 3 s on average (residual variance, $\sigma_{ip} = 3.16$ s, $p < .001$). Items varied significantly in their solution times (item solution time intercept standard deviation, $\tau_{0i} = 2.46$ s, $p < .001$). The use of min-counting strategies varied in its effect on item solution times (item solution time slope standard deviation, $\tau_{1i} = 11.58$ s, $p < .001$). Solution times and the effects of min-counting strategies on solution times were again significantly correlated at a similar magnitude, $r = .79$. In other words, items that took longer to solve overall tend to take even longer if min-counting strategies were used.

People also varied significantly in their solution times (person solution time intercept standard deviation, $\tau_{0p} = 1.91$ s, $p < .001$), but on average people took approximately 4 s to solve problems (person solution time grand mean, $\gamma_{00} = 4.05$, $p < .001$). The overall effect of min-counting was an increase in solution times of about 5 s (person solution time slope mean, $\gamma_{10} = 4.86$, $p < .001$). However, the use of min-counting strategies varied significantly in its effect on people's solution times (person solution time slope standard deviation, $\tau_{1p} = 2.57$ s, $p < .001$). Again, the significant covariance between the random intercept and slope, $r = -.38$, $p = .001$ indicated that people who were slower overall tended to have a lower solution time penalty for use of counting, but faster people tended to take longer if they used min-counting strategies.

Items varied significantly in their predicted accuracy (i.e., their difficulty; item accuracy intercept variance, $\tau^2_{0i} = .26$, $p < .001$). The use of min-counting strategies varied in its effect on item accuracy (item accuracy slope variance, $\tau^2_{1i} = .19$, $p < .001$). Again, item difficulty and the effects of min-counting strategies on item accuracy were not significantly related, $r = -.27$, $p = .19$.

People also varied significantly in their accuracy (person accuracy intercept variance, $\tau^2_{0p} = .22$, $p < .001$), but the average person had a probability of .97 of correctly solving the average item (person accuracy threshold, $\gamma_{00} = -1.91$, $p < .001$). Overall, for the average person, the use of min-counting strategies did not significantly impact the likelihood of generating a correct answer (person accuracy slope mean, $\gamma_{10} = -.19$, $p = .09$). However, people differed significantly in how accurately they used min-counting (person accuracy slope variance, $\tau^2_{1p} = .23$, $p < .001$). Unlike in the first random sample, the covariance between the random intercept and slope, $r = -.42$, $p = .03$ was significant, indicating that people who were less accurate overall were even less accurate when they used min counting.

Received April 26, 2017
Revision received May 9, 2018
Accepted June 11, 2018 ■