Journal of School Psychology 97 (2023) 77–100

Contents lists available at ScienceDirect

Journal of School Psychology

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

Ordinal models to analyze strategy sophistication: Evidence from a learning trajectory efficacy study

T.S. Kutaka [a,*,1], P. Chernyavskiy [b,1], J. Sarama [c], D.H. Clements [c]

[a] University of Virginia, Center for the Advanced Study of Teaching and Learning, Ridley 236, PO Box 800784, Charlottesville, VA 22904, United States of America
[b] University of Virginia, Department of Public Health Sciences, PO Box 800717, Charlottesville, VA 22908, United States of America
[c] University of Denver, Morgridge College of Education, Marsico Institute for Early Learning, 1999 East Evans Avenue, Denver, CO 80208, United States of America

ARTICLE INFO

Editor: Craig A. Albers
Action Editor: Nick Benson

Keywords:
Problem-solving strategies
Strategy sophistication
Mathematics learning trajectories
Early childhood mathematics education
Research methods

ABSTRACT

Investigators often rely on the proportion of correct responses in an assessment when describing the impact of early mathematics interventions on child outcomes. Here, we propose a shift in focus to the relative sophistication of problem-solving strategies and offer methodological guidance to researchers interested in working with strategies. We leverage data from a randomized teaching experiment with a kindergarten sample whose details are outlined in Clements et al. (2020). First, we describe our problem-solving strategy data, including how strategies were coded in ways that are amenable to analysis. Second, we explore what kinds of ordinal statistical models best fit the nature of arithmetic strategies, describe what each model implies about problem-solving behavior, and how to interpret model parameters. Third, we discuss the effect of "treatment", operationalized as instruction aligned with an arithmetic Learning Trajectory (LT). We show that arithmetic strategy development is best described as a sequential stepwise process and that children who receive LT instruction use more sophisticated strategies at post-assessment, relative to their peers in a teach-to-target skill condition. We introduce latent strategy sophistication as an analogous metric to traditional Rasch factor scores and demonstrate a moderate correlation them ($r = 0.58$). Our work suggests strategy sophistication carries information that is unique from, but complimentary to traditional correctness-based Rasch scores, motivating its expanded use in intervention studies.

Early mathematics competencies are the strongest predictors of later reading and math outcomes, even after accounting for preschool cognition, attention, socioemotional skills, as well as individual- and family-level characteristics (Claessens & Engel, 2013; Duncan et al., 2007). Numbers and operations knowledge is a particularly critical instructional goal in the early years as it can predict mathematics difficulties (Garon-Carrier et al., 2018; Geary, 2011; Jordan et al., 2009, 2010; VanDerHeyden et al., 2011) and later mathematics success (Claessens et al., 2009; Duncan et al., 2007; Pagani et al., 2010; Watts et al., 2014). To strengthen preschool and early primary mathematics outcomes, early learning and care programs launched federal-, state-, and foundation-supported initiatives in the form of direct (e.g., early screening for math difficulties; response to intervention models) and indirect (e.g., teacher professional

* Corresponding author.
E-mail addresses: traci.kutaka@virginia.edu (T.S. Kutaka), pchern@virginia.edu (P. Chernyavskiy), julie.sarama@du.edu (J. Sarama), douglas.clements@du.edu (D.H. Clements).
[1] Indicates equal contribution.

https://doi.org/10.1016/j.jsp.2023.01.002
Received 1 February 2022; Received in revised form 10 September 2022; Accepted 11 January 2023
Available online 7 February 2023

development; school-family partnerships) interventions. These programs typically record progress towards achieving fluency, which serves as an important metric for constructing learning efficacy and effectiveness arguments.

The ways in which researchers and evaluators conceptualize mathematics learning carry consequences for programmatic funding and investment, as well as the metrics in which we place our trust (Rodrigues, 2021; VanDerHeyden & Harvey, 2013). We contend that analysis of problem-solving strategy sophistication enables us to systematically collect artifacts of children's mathematical thinking not fully captured through accuracy or fluency alone. Indeed, fluency is built on the dexterity and mastery children hold over numeracy concepts as they move through the problem-solving process, as well as encounter and work though different types of numerical tasks (Baroody, 2003, 2006).

To this end, we introduce the process of how arithmetic strategy sophistication is coded and then quantified on an ordinal scale using data from a randomized teaching experiment. We apply Bayesian modeling techniques to three important sources of variation, including (a) the demands of items, (b) individual students' pre-assessment arithmetic competences, and (c) opportunities to learn (defined by child-level assignment to experimental condition). Through our methods, we demonstrate that analysis of item, student, and experimental effects offer researchers and evaluators nuanced portraits of student sense-making uniquely expressed through examination of strategy sophistication data.

1. Centralizing problem-solving strategies as an outcome of interest in early mathematics education research and evaluation

Young children have nascent, intuitive mathematical ideas (Clements & Sarama, 2021; Ginsburg et al., 2008) that are usually not revealed by correct responses alone. We can uncover what children know and can do mathematically by analyzing the concepts and processes children use as they attempt to solve problems. These processes, including problem-solving strategies, have been placed on par with content standards as goals of mathematics education (National Council of Teachers of Mathematics [NCTM], 2000; National Governors Association & Council of Chief State School Officers, 2010; National Research Council [NRC], 2001, 2009). For example, "Making sense of problems and persevering in solving them" is the first mathematical principle in the Common Core, whereas problem-solving serves as one of NCTM's Process Standards. In this view, problem-solving strategy development is both a mathematical tool, as well as a desired mathematics outcome, and ought to be part of how we construct arguments about learning efficacy and effectiveness.

1.1. Research on strategy development in early childhood mathematics

Early studies in cognitive development lay the foundation for how we define and study strategies: a problem-solving strategy is a procedure that is goal-directed, selective, and intentional (Flavell, 1970; Naus & Ornstein, 1983; Siegler & Jenkins, 1989). This definition of strategy does not require planning, but rather only the intention of achieving a goal.

Experimental studies have demonstrated that variability is a central characteristic of the learner's cognitive system (as opposed to an artifact of measurement or random error). Siegler (1991) used microgenetic methods to describe how young children develop, refine, and generalize strategies within the context of arithmetic problems. This work culminated in the overlapping waves theory, which posits that children's concepts are not static: children can think about and solve the same problem using multiple strategies. In fact, preschool and early primary school-aged children were observed to solve a single digit addition problem via three distinct strategies (Siegler, 1987; Siegler & Jenkins, 1989). Another reported source of variation is primary-grade gender differences for strategy preference when solving problems in the domain of numbers and operations (Carr & Alexeev, 2011; Carr & Davis, 2001; Fennema et al., 1998; Supovitz et al., 2021; Zhu, 2007).

Within mathematics education research, a problem-solving strategy is conceptualized as "engaging in a task for which the solution method is not known in advance. To find a solution, students must draw on their knowledge, and through the process, they will often develop new mathematical understanding" (NCTM, 2000, p. 52). The development of increasingly sophisticated strategies is not only a valued outcome in itself, but also because they are foundational for, and harbingers of, children's development of correct calculations and fact fluency (Clements & Sarama, 2021). For example, a conceptually based intervention that developed children's counting and reason strategies was more efficacious and more successful than either regular classroom instruction or a direct instruction approach in promoting progress towards fluency and fluently itself (Baroody, 2016).

1.2. Using learning trajectories to quantify problem-solving strategy sophistication as a useful metric for child learning

There are several documented programs explicitly designed to support problem-solving strategies. These approaches include (a) teaching general phases of problem-solving, such as understanding the problem, devising a plan, carrying out the plan, then checking and interpreting (Cai & Brook, 2006; Polya, 1985); (b) providing problems in a developmental sequence, from simple "result unknown" to more complex "start unknown" or "comparison" problem types (Carpenter et al., 2015); and (c) simultaneously placing an emphasis on skills, content, and problem-solving processes (Lesh et al., 2000; National Mathematics Advisory Panel, 2008). Recent work on learning trajectories synthesizes these approaches (Clements & Sarama, 2021; Daro et al., 2011; Lobato & Walter, 2017; Simon, 1995) and situates problem-solving strategies within developmental progressions, which serves as the theoretical and instructional basis for the present study.

Learning trajectories (Clements & Sarama, 2021) foreground student cognition. Policymakers, educators, and researchers deem them to be useful tools for guiding teaching and learning standards, cycles of planning and instruction, professional development, and assessment (Frye et al., 2013; NRC, 2009). Broadly defined, a mathematical learning trajectory builds on empirical research to describe

probable pathways of learning over time within a particular topic (NRC, 2009). Indeed, a growing body of research describes the contributions of learning trajectories to teachers' professional knowledge and skills for teaching (Kutaka et al., 2018; Sarama & Clements, 2019; Supovitz et al., 2018), as well as student learning (Clements et al., 2020; Kutaka et al., 2017; Sarama et al., 2021).

However, what has received less attention is how learning trajectories produce empirically grounded descriptions of student problem-solving strategies, which also enable us to map and rank strategies on a sophistication scale. Problem-solving strategies reflect the implicit or explicit awareness of the principles that govern a body of knowledge within a domain (Rittle-Johnson et al., 2001); one example of this is a child explicitly or intuitively knowing that $3 + 4 = 4 + 3$, as changing the order of the operands does not change the result (i.e., commutative property). Increasing the sophistication of problem-solving strategies becomes necessary as children encounter increasingly complex problems and concepts (Carpenter et al., 1998; Maloney et al., 2014).

Consequently, the purpose of the present study was to examine whether instruction aligned with the learning trajectories (Clements & Sarama, 2021) can support the development of more sophisticated arithmetic problem-solving strategies. We leverage data from a recent teaching experiment conducted in a Mountain West state within an urban school district. An in-depth description of the experimental design, content of the instructional sessions, fidelity of implementation, and theoretical rationale for the construction of the comparison group for this randomized efficacy experiment can be found in Clements et al. (2020). The validation and psychometric functioning of the items used for Clements et al. (2020) and the present study permits us to accept the ordering of strategies from the least to the most sophisticated (see Table 1 and the Instruments section). However, given the dearth of studies that center strategy development as the outcome of interest in educational efficacy and effectiveness research, there is little guidance on how to analyze strategy sophistication data in the literature.

1.3. Present study

Within the present study, we posit that learning trajectories serve as a powerful instructional platform that can increase the sophistication of arithmetic problem-solving strategies. Additionally, learning trajectories enable the coding of strategies according to their relative sophistication, making them amenable to analysis. Thus, our goal was to produce guidance for education researchers and evaluators who work with strategy sophistication as the outcome of interest.

We carried out this work in three phases, which is described in more detail within the Method section below. First, we describe our problem-solving strategy data, including how arithmetic strategies were coded. Second, we explore what kinds of statistical models for ordinal data best fit the nature of arithmetic strategies, describe what each model implies about problem-solving behavior, and offer guidance for how to interpret model parameters. Finally, we discuss the effect of "treatment" (receiving learning trajectories-aligned instruction) given the ordinal nature of strategy data, as well as compare conditional treatment effects (that apply to the "typical" item of median difficulty) to marginal treatment effects ("averaged over the population of items") in the context of a teaching experiment.

Table 1
Description of arithmetic strategies and coding on ordinal sophistication scale.

Strategy	(Brief) Operational Definition	Example	Coded Level of Sophistication
Guessing	Solution is not grounded in a firm understanding of the problem.	5 + 2: "I don't know–10?"	1 (L1)
Trial & Error	Estimated number is tried and adjusted iteratively.	6 + _ = 13 "6 plus 5 is [counts out objects, sets of 6 and 5] 11…umm [adds one, recounts] 12…one more, so. [counts added set]…7.	1 (L1)
Counting All	Produce sets of objects representing each number, combine or separate, count the result.	5 + 2: make set of 5 items, count out 2 more items, and count all those starting again at "one." If no counting errors–report "7."	2 (L2)
Counting On	Curtail counting by assuming the initial quantity, using objects to keep track of second quantity as the counting act proceeds.	5 + 2: "Fiiiive… [puts up one finger], six [puts up another finger], seven [recognizes finger pattern as "2″]: Seven!"	3 (L3)
Counting On – Abstract	Assume the initial quantity, using subitized patterns or "double-counting" to keep track of the second quantity as the counting act proceeds.	5 + 2: "Fiiiive… [counts a rhythmic pattern of two] *six, seven.* Seven!"	3 (L3)
Jump Strategy	Starting with one number, move along a mental number line by tens and ones.	28 + 31: "28, 38, 48, 58, and one more is 59."	4 (L4)
Combination	Uses retrieval.	5 + 2: "Seven."	4 (L4)
Derived Combination	Uses a known combination, to figure out another combination.	6 + _ = 13: "6 plus 6 is 12, so…seven!"	4 (L4)
Decomposition	Decomposes numbers and recomposes new parts.	9 + 6: "9 and 1 is 10; 10 and 5 is 15." 28 + 35: "20 and 30 is 50, 8 and 5 is 13; 50 and 13 is 63."	4 (L4)

2. Method

2.1. Current dataset

The data for this study come from one of a series of twelve teaching experiments testing the theoretical assumptions of learning trajectories across multiple early mathematics topic strands (Institute for Education Sciences Grant #R305A150243). The current dataset contains information we used to compare the effects of two instructional treatments on kindergartners' arithmetic thinking and reasoning. One experimental condition was designed to move children through consecutive levels of a research-based learning trajectory (i.e., LT condition; $n = 143$; see Clements & Sarama, 2021, or LT^2 website [https://www.learningtrajectories.org/learning_trajectories] for a description of the levels of thinking that compose the trajectory). LT-aligned instructional decisions were defined by administering activities at $N + 1$ for each child, where N was defined by the child's current level of arithmetic thinking and reasoning. The levels of the LT offered guidance about what kinds of arithmetic problems children were ready to encounter and work through. We include an example of an LT activity in Appendix A.

The counterfactual condition provided similar activities but focused directly on the target skills, thereby skipping levels of the trajectory between each child's initial level of thinking and the target level (i.e., SKIP condition; $n = 148$). Both treatments included up to 20 one-on-one sessions, where instructors emphasized multiple methods and mediums for representing the agents, actions, and relationships in story problems. Further details regarding the implementation of this study are described in Clements et al. (2021).

We provide a graphical overview of post-assessment sophistication ratings by treatment condition (i.e., LT vs. SKIP) and (a) pre-

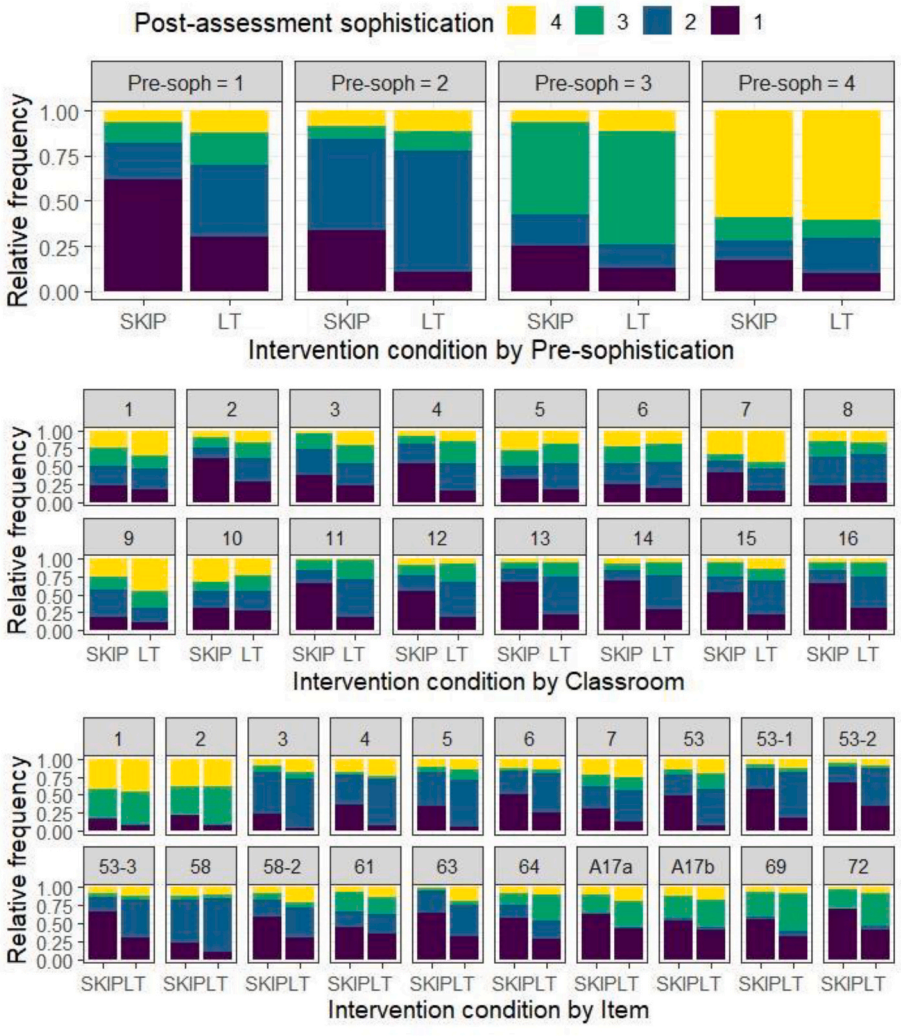


Fig. 1. Observed frequencies of post-assessment sophistication ratings by treatment condition (LT versus SKIP) and pre-sophistication (Top Panels), classroom (Middle Panels), and item (Bottom Panels).
Note. Item panels are arranged in order of their developmental progression from easiest (Item "1") to most difficult (Item "72").

assessment sophistication (Fig. 1 top set of panels), (b) classroom (Fig. 1 middle set of panels), and (c) item (Fig. 1 bottom set of panels). We note three trends. First, students tended to use the same strategy at post as they did at pre-assessment, except for students who used L1 strategies at pre-assessment. That said, students in the LT condition tended to use more sophisticated strategies than students in the SKIP condition in all but the highest level of pre-sophistication. Second, there was marked variability in strategy sophistication by classroom. Third, students tended to use more sophisticated strategies on easier items (e.g., Item 1, Item 2) compared to more difficult items (Item A17a through 72).

2.2. Participants

In the original experiment, we acquired consent from 319 kindergarten students from 16 classrooms within four schools in an urban school district in a Mountain West state. Twenty-eight children dropped out of the study. Six of those students moved to a different school in the middle of the study. The remaining 22 of these students demonstrated behaviors challenging for adults and did not participate in the experiment at the request of their classroom teachers. Thus, 291 students compose the analytic sample used here. The differential attrition rate was found to be statistically non-significant, as reported in Clements et al. (2020).

2.3. Measures

The outcome of interest included the strategies used on 20 items from the Research-based Early Mathematics Assessment (REMA; Clements et al., 2008) and the Test of Early Mathematics Ability – 3rd edition (TEMA-3; Ginsburg & Baroody, 2003). The REMA and TEMA-3 are measures of pre-K through Grade 3 children's mathematical knowledge. Noteworthy is that all items are story problems; no worksheets listing equations were used (although blank paper and pencils were available to them as problem-solving tools). Items were administered in individual interviews of each child, with explicit protocol, coding, and scoring procedures. Scoring incorporates both correctness of responses and strategies used by children. All student assessments were video recorded.

Unidimensionality for this collection of items was established through PFA and Cronbach's $\alpha = 0.85$. Thus, Rasch scores were constructed and information, an analog of reliability, was 0.80 across the latent continuum.

2.3.1. Strategy sophistication coding for the arithmetic assessment

Included as part of the assessment record was a list of common strategies deployed by young children for every item of the assessment as developed through the REMA (which has evidence of validation, as reported in the section above) and grounded in Clements and Sarama's learning trajectories. These empirically validated strategies were grouped and ordered by sophistication on a 4-point ordinal scale; examples include the following: 1 = *trial and error*; 2 = *count all*; 3 = *count on*; and 4 = *known combinations and decomposition* (see Table 1). Appendix A contains a sample item alongside the strategy codes. Children were not informed whether they answered each item correctly during the assessment; thus, their strategy selection was unaffected by feedback.

To generate data for this study (IES Grant #R305A200100), a team of two coders, who also served as members of the instructional team, revisited the assessment videos to review the strategy codes. The coding team was trained and then the team established reliability against the master codes of the project lead. Each member of the coding team achieved 91% agreement across 10% of the videos, stratified by experimental condition and treatment site. In total, we collected strategy sophistication ratings, pre- and post-assessment, from 291 children (randomized such that 143 were in the LT condition and 148 were in the SKIP condition), across 16 kindergarten classrooms within four Title I schools located in a Mountain West state of the United States. The observed frequencies for pre- and post-assessment are in Table 2.

Table 2
Frequency table of post-assessment strategy sophistication by treatment condition and pre-assessment.

Strategy Sophistication						
Pre-Sophistication Level	Post-Sophistication Level					
	1	2	3	4	NA	Total
	SKIP					
1	743 (51.56%)	247 (17.14%)	133 (9.23%)	80 (5.56%)	238 (16.52%)	1441 (100%)
2	124 (30.85%)	186 (46.27%)	27 (6.72%)	30 (7.46%)	35 (8.71%)	402 (100%)
3	75 (22.73%)	53 (16.06%)	157 (47.58%)	19 (5.76%)	26 (7.88%)	330 (100%)
4	48 (15.74%)	30 (9.84%)	37 (12.13%)	166 (54.43%)	24 (7.87%)	305 (100%)
NA	219 (45.44%)	83 (17.22%)	41 (8.51%)	41 (8.51%)	98 (20.33%)	482 (100%)
Total	1209 (40.84%)	599 (20.24%)	395 (13.34%)	336 (11.35%)	421 (14.22%)	2960 (100%)
	LT					
1	364 (25.72%)	472 (33.36%)	219 (15.48%)	146 (10.32%)	214 (15.12%)	1415 (100%)
2	42 (9.79%)	264 (61.54%)	42 (9.79%)	46 (10.72%)	35 (8.16%)	429 (100%)
3	39 (11.75%)	41 (12.35%)	192 (57.83%)	36 (10.84%)	24 (7.23%)	332 (100%)
4	30 (9.17%)	62 (18.96%)	30 (9.17%)	188 (57.49%)	17 (5.20%)	327 (100%)
NA	64 (17.93%)	139 (38.94%)	65 (18.21%)	30 (8.40%)	59 (16.53%)	357 (100%)
Total	539 (18.85%)	978 (34.20%)	548 (19.16%)	446 (15.60%)	349 (12.20%)	2860 (100%)

2.4. Analytical plan

To accurately model the nature of ordinal strategy data (ordered by sophistication), we describe three families of statistical models and subsequently detail their estimation through a Bayesian paradigm. Bayesian analyses are becoming more common in the social sciences; to this end, we provide a brief primer on how to apply these methods and select reasonable prior distributions in the context of education research. Here, we use the modern Hamiltonian Monte Carlo for estimation and describe the relevant computational details in Appendix B.

2.4.1. Statistical models for ordinal data

Let Y denote the column vector of post-assessment strategy sophistication ratings collected on $i = 1, \ldots, I$ items from $j = 1, \ldots, J$ students within $k = 1, \ldots, K$ classrooms. Sophistication ratings were comprised of $c = 1, \ldots, C$ ordered (ordinal) sophistication categories, such that $Y_{ijk} = 1$ denotes the lowest level of sophistication and $Y_{ijk} = C$ denotes the highest level of sophistication, where $C = 4$ for the current data. Models contain $C - 1$ logit (i.e., log-odds) terms (Table 3) and thus require the estimation of $C - 1$ "threshold" parameters; for the current data: $\Theta = (\theta_1, \theta_2, \theta_3)$. What follows is a description of how these logit terms are constructed, how covariates influence the logit terms, and what each model implies for the strategy sophistication data-generating process. The R code for this analysis can be found in GitHub [www.github.com/pchernya/Strat_sophist_ordinal_models].

In its most general form, using the subscript c to reflect possible response category-specific effects, let the following be the linear predictor on the logit scale:

$$\eta_{ijkc} = x_{ijk}^{'}\beta_c + u_{ic} + v_{jc} + w_{kc},$$

where x_{ijk}' contains treatment effects and pre-assessment strategy sophistication ratings, β_c is the associated vector of unknown coefficients, u_{ic} are item random intercepts, v_{jc} are student random intercepts, and w_{kc} are class random intercepts. Item random effects are crossed with both student and class random effects, whereas student effects are nested within class effects.

We denote $\pi_{ijkc} = P(Y_{ijk} = c)$ as the probability that the *jth* student in the *kth* classroom employs strategy rating c on item i. Conditional on the parameters, the likelihood for the full response vector Y is defined as:

$$L(\mathbf{Y}|\boldsymbol{\Theta}, \boldsymbol{\beta}, \mathbf{u}, \mathbf{v}, \mathbf{w}) = \prod_{i=1}\prod_{j=1}\prod_{k=1}\prod_{c=1}(\pi_{ijkc})^{d_{ijkc}},$$

where $d_{ijkc} = 1$ if $Y_{ijk} = c$ and $d_{ijkc} = 0$ otherwise, and $\sum_C \pi_{ijkc} = 1$. The three families of models examined in this article construct π_{ijkc} in different ways; however, as shown by the equation above, the likelihood formed by these probabilities remains identical.

2.4.1.1. Cumulative model. The cumulative logit ordinal model (e.g., McCullagh, 1980) is among the most widely used methods to analyze ordinal categorical responses. This model is motivated by a process where a continuous latent variable is discretized into C ordered categories, using $C - 1$ cut points or "thresholds". In the context of early childhood education, this latent variable is typically called "ability" (e.g., Council, 2000; Kutaka et al., 2018), which forms the basis of several common item-response models, such as the Rasch Model (e.g., Andrich, 1978) and the Graded Response Model (e.g., Samejima et al., 1997).

In application to strategy sophistication, the cumulative model assumes there exists a single latent "sophistication continuum", which we observe only through a finite number of problem-solving behaviors mapped into ordered sophistication categories. Under the cumulative logit model, the linear predictor is incorporated with each logit term as follows:

$$log\left(\frac{P(Y_{ijk} \le c)}{P(Y_{ijk} > c)}\right) = \theta_c - \eta_{ijk}$$

and the probabilities of interest are computed as:

$$\pi_{ijkc} = P(Y_{ijk} \le c+1) - P(Y_{ijk} \le c) = logit^{-1}(\theta_{c+1} - \eta_{ijk}) - logit^{-1}(\theta_c - \eta_{ijk}).$$

Importantly, linear predictor η_{ijk} does not depend on response category c and the threshold parameters are assumed to be ordered such that $-\infty < \theta_1 \le \theta_2 \le \theta_3 < \infty$. The constraint in which only the thresholds θ_c depend on response categories is typically called the "proportional odds" assumption. When random effects are present, this assumption is conditional on the random effects: after

Table 3
Construction of logit terms for three families of models for ordinal data.

Logit	Sequential	Adjacent Category	Cumulative
*L*1	*P*(Y = 1) vs. *P*(Y ≥ 2)	*P*(Y = 1) vs. *P*(Y = 2)	*P*(Y = 1) vs. *P*(Y ≥ 2)
*L*2	*P*(Y = 2) vs. *P*(Y ≥ 3)	*P*(Y = 2) vs. *P*(Y = 3)	*P*(Y ≤ 2) vs. *P*(Y ≥ 3)
*L*3	*P*(Y = 3) vs. *P*(Y = 4)	*P*(Y = 3) vs. *P*(Y = 4)	*P*(Y ≤ 3) vs. *P*(Y = 4)

Note. Current strategy data have four ordinal sophistication categories, where $Y = 1$ denotes the lowest level of sophistication and $Y = 4$ denotes the highest level of sophistication. The Sequential logit terms can be equivalently expressed as: $L_1 = P(Y = 1)$*vs.* $P(Y > 1)$, $L_2 = P(Y = 2)$*vs.* $P(Y > 2)$, $L_3 = P(Y = 3)$*vs.* $P(Y > 3)$.

accounting for variability due to item, student, and classroom, the effect of treatment is independent of strategy sophistication level. This can be an unrealistic assumption because we often expect treatment to have greater impact at lower levels of sophistication, which is consistent with other mathematics intervention research (e.g., Clements et al., 2013).

Nevertheless, cumulative logit models offer the most parsimony relative to other models for ordinal data, offer the most straightforward interpretations, and may serve as useful approximations of the true data-generating process even when the proportional odds assumption is violated. In addition, cumulative logit models are robust to different strategy coding choices that may vary across disciplines and mathematical tasks, thus producing different sophistication scales. For example, cumulative model parameter estimates are unaffected by the total number of sophistication (i.e., response) categories, which enables informative comparison across studies (e.g., Agresti, 2018).

2.4.1.2. Adjacent category model. The adjacent category model considers pairs of adjacent responses (c vs. $c+1$), as opposed to the full response scale, as in the cumulative model. Adjacent category models are more general than cumulative models because they do not require thresholds θ_c to be ordered and allow different parameter estimates for different response categories, thereby relaxing the proportional odds assumption. The linear predictor is assigned to the logit term as:

$$log\left(\frac{P(Y_{ijk}=c)}{P(Y_{ijk}=c+1)}\right)=\theta_c-\eta_{ijkc},$$

where probabilities of interest under the adjacent category model are defined as:

$$\pi_{ijkc}=\frac{exp(\theta_c-\eta_{ijkc})}{1+\sum_{l=1}^{C-1}exp(\theta_l-\eta_{ijkl})}.$$

In addition to the thresholds, all or some parameters in the linear predictor η_{ijkc} may depend on response category c. Importantly, the adjacent category model allows treatment effects to be response category-specific; for example, the greatest impact may occur for strategies at the lower end of the sophistication spectrum. When adjacent category models are parameterized as shown above, parameter estimates >0 increase the odds of observing the next-highest level of sophistication ($c+1$) vs. the current level of sophistication (c), and vice versa.

Item, student, and classroom random intercepts may also vary with response category, which produces $C-1=3$ sets of effects per random intercept. These three sets of effects are expected to share a considerable amount of information and are thus allowed to be correlated. Joint distributions are assumed to be mean-zero Multivariate Normal with covariance matrices:

$$\boldsymbol{\Sigma}_u=\begin{pmatrix}\sigma_{u1}&0&0\\0&\sigma_{u2}&0\\0&0&\sigma_{u3}\end{pmatrix}\boldsymbol{\Omega}_u\begin{pmatrix}\sigma_{u1}&0&0\\0&\sigma_{u2}&0\\0&0&\sigma_{u3}\end{pmatrix},$$

$$\boldsymbol{\Sigma}_v=\begin{pmatrix}\sigma_{v1}&0&0\\0&\sigma_{v2}&0\\0&0&\sigma_{v3}\end{pmatrix}\boldsymbol{\Omega}_v\begin{pmatrix}\sigma_{v1}&0&0\\0&\sigma_{v2}&0\\0&0&\sigma_{v3}\end{pmatrix},$$

$$\boldsymbol{\Sigma}_w=\begin{pmatrix}\sigma_{w1}&0&0\\0&\sigma_{w2}&0\\0&0&\sigma_{w3}\end{pmatrix}\boldsymbol{\Omega}_w\begin{pmatrix}\sigma_{w1}&0&0\\0&\sigma_{w2}&0\\0&0&\sigma_{w3}\end{pmatrix}$$

for item, student, and classroom random intercepts, respectively. Standard deviations σ_{uc}, σ_{vc}, σ_{wc} measure the response category-specific variability. Accordingly, $\boldsymbol{\Omega}_u$, $\boldsymbol{\Omega}_v$ and $\boldsymbol{\Omega}_w$ are the 3 × 3 correlation matrices (with elements $\rho_{u12}, \rho_{u13}, \ldots, \rho_{v12}, \rho_{v13}, \ldots \rho_{w23}$) that capture correlation between the category-specific item, student, and classroom effects. Correlations between category-specific random effects are sensible because a classroom with an advantaged group of students, for example, may have a positive random intercept for all 3 category-specific logits, inducing a positive correlation.

Category-specific item effects account for story problem characteristics (e.g., largest number used, mathematics operation required) that encourage the use of some strategies over others. Category-specific student effects reflect latent personal preferences or tendencies to favor a particular set of strategies. Given that there was no feedback provided during pre- and post-assessment, these preferences are independent of whether students answered the items correctly. In the context of strategy sophistication, we refer to these as "latent sophistication", analogous to the nomenclature of "latent ability" in early childhood mathematics. Category-specific classroom effects measure how latent strategy preferences vary by classroom. These effects likely reflect the impact of unmeasured socio-demographic variables because the 16 classrooms in our data were situated across four school districts that admit students from different neighborhoods of the same metropolitan area.

Although adjacent category models are more flexible than cumulative models, Bürkner and Vuorre (2019) argued that they do not arise from an interpretable data-generating process. Peyhardi et al. (2015) showed that adjacent category models are not motivated by an underlying latent variable, and Fullerton and Xu (2016, p. 39) noted that choice between cumulative and adjacent category models can be driven by whether interest lies in the latent variable - the "sophistication continuum" in our data - or specific pairs of observed response categories. We favor models that enable the study of the underlying continuum, because observed sophistication categories

undoubtedly vary with the mathematical task, specific coding principles, and may be subject to coding errors. Of course, if estimated category-specific effects do not vary substantially between response categories, the cumulative model may be selected over the adjacent category model via an information criterion, such as the Watanabe-Akaike Information Criterion (WAIC; Watanabe, 2013) or the Leave-One-Out Information Criterion (LOOIC; Vehtari et al., 2017). Here, we investigate several competing models, each with an increasing number of category-specific effects, which in turn further relax the parallel odds assumption inherent to the cumulative model (see Table 3).

2.4.1.3. Sequential model. If the data-generating process is inherently sequential, such that subjects "experience" response categories in a specific order, the sequential logit family of models (e.g., Tutz, 1991, 2005) may be more appropriate than either cumulative or adjacent category models. Contrasted against a single sophistication continuum (cumulative model) or lack thereof (adjacent model), the sequential model motivates us to consider the odds of making a latent "transition" from lower-sophistication strategies to higher-sophistication strategies. Ideas behind sequential logit models have long been the basis of Partial Credit Item Response models (Masters, 1982), which are often used in psychometric analysis of mathematical ability. In these analyses, when the correct response is provided, the sequential model *infers* that the student has successfully completed all steps necessary to arrive at the final answer, although only the outcome of the last step (i.e., the final answer) is recorded.

To draw parallels to strategy sophistication, under the sequential model, we only observe the result (sophistication rating) of the last latent "transition" a student makes and infer that they completed all transitions before it. Thus, the model treats growth in sophistication in a stepwise manner: the use of the most sophisticated strategy is predicated on the mastery of all strategies that are less sophisticated. Furthermore, threshold parameters θ_c now govern each of the latent transitions from one sophistication level to another and no longer represent cut-points of a single continuum. Under the sequential model there exist $C-1=3$ latent continuums, where θ_1 governs transitions from $Y=1$ to $Y\geq 2$, θ_2 governs transitions from $Y=2$ to $Y\geq 3$ among those who reached $Y=2$, and θ_3 the transitions from $Y=3$ to $Y=4$, for those who reached $Y=3$.

Nomenclature for sequential models has been inconsistent and it depends, for example, on whether we conceptualize the process as "stepping up" vs. "stepping down", and whether we are interested in "stopping at" response category c vs. "continuing beyond" c (Smithson & Merkle, 2013, p. 92). For strategy sophistication, we focus on "stopping ratio" models - also called "continuation-ratio" models by Agresti (2018, p. 191) - that contrast $P(Y=c)$ vs. $P(Y>c)$. In the context of this article, we are modeling whether sophistication stops at some level of sophistication ($Y=c$) vs. sophistication continues to higher levels ($Y>c$). Furthermore, we feel it is not logical to conceptualize changes in sophistication as a "stepping down" process, so we conceptualize strategy sophistication as starting from $Y=1$ and "stepping up" to $Y>1$.

In a sequential model, covariates are assigned to the logit term:

$$log\left(\frac{P(Y_{ijk}=c)}{P(Y_{ijk}>c)}\right)=\theta_c-\eta_{ijkc}$$

with probabilities of interest defined by:

$$\pi_{ijk1}=P(Y_{ijk}=1)=logit^{-1}(\theta_1-\eta_{ijk1}),$$

$$\pi_{ijkc}=P(Y_{ijk}=c)=logit^{-1}(\theta_1-\eta_{ijk1})\times\prod_{c=2}[1-logit^{-1}(\theta_c-\eta_{ijkc})\,],\text{ for } c=2,\ldots,C-1,$$

$$\pi_{ijkC}=P(Y_{ijk}=C)=1-logit^{-1}(\theta_C-\eta_{ijkC}).$$

Reflecting the sequential nature of the data, these probabilities are the likelihood of successfully transitioning through each lower-ranked sophistication category, but failing to clear the threshold for the next-highest sophistication (Bürkner & Vuorre, 2019; Tutz & Groll, 2013).

As in adjacent category models, the linear predictor for sequential models may depend on response category c, and thus may include category-specific treatment effects and random intercepts for items, students, and classrooms as previously described. We also bring attention to the $-$ sign in front of the linear predictor; parameterized in this manner, any coefficient >0 increases the odds of transitioning (continuing) to higher levels of sophistication rather than remaining at what is currently attained. Conversely, coefficients <0 indicate greater odds of remaining (stopping) at the level of sophistication that is currently attained. Furthermore, thresholds θ_1, θ_2, θ_3 are not assumed to be ordered, which reflects potentially different difficulties of transitioning, for example, from $Y=1$ to $Y\geq 2$ for *all* children vs. $Y=3$ to $Y=4$ among those who have reached $Y=3$. The former may prove to be more difficult because the latter transition occurs *conditional* on a moderate level of arithmetic skill being present (i.e., $Y=3$ is assumed to be attained for the transition to $Y=4$ to occur).

2.4.2. Bayesian estimation

2.4.2.1. Primer on Bayesian analysis. Bayesian analysis necessitates careful specification of *prior distributions*, which encode the researchers' understanding about parameters (in general, $\alpha_1, \ldots, \alpha_p$) before observed data ($Y$) enters the model. By rearranging terms in the classic Bayes Theorem, we can show that the *posterior distribution* $p(\alpha_1,\ldots,\alpha_p|\,Y)$, distribution of model parameters conditional on

the data, is proportional to:

$$p(\alpha_1, \ldots, \alpha_p|Y) \propto f(Y|\alpha_1, \ldots, \alpha_p) \times p(\alpha_1) \times \ldots \times p(\alpha_p),$$

which is the product of the *likelihood* $f(Y|\alpha_1, \ldots, \alpha_p)$ and each *prior distribution* $p(\alpha_1) \times \ldots \times p(\alpha_p)$. When the number of observations is relatively small, as is the case with small- and moderate-scale teaching experiments, prior distributions carry as much or more weight in computing the posterior, as does the likelihood. Conversely, as the number of observations grows, the likelihood carries increasingly more weight and is said to "dominate the priors". In other words, with enough data, the exact prior specifications may become unimportant. However, we caution the reader that even in data-rich scenarios, the amount of information to estimate some model parameters may be scarce, and so the priors for those parameters may still be influential. This often happens with parameters that operate on the latent scale, for example: random effect standard deviations and the corresponding correlations/covariances.

2.4.2.2. Prior Selection selection in Bayesian analysis. Researchers select a statistical model with a particular estimation task in mind; thus, it is not useful to consider priors outside the context of the data to which the model is applied (Gelman et al., 2017; Gelman & Hennig, 2017). Consider, for example, the prior distribution for a categorical treatment effect (β_{TRT}) in a cumulative logit model within the context of a teaching experiment. Under a typical "uninformative" Normal(0,100) prior, the researchers presume—before collecting any data—that there is no treatment effect on average, and the range of plausible treatment effects (i.e., the 95% interval) *on the latent scale* is $(-196, 196)$. This range is absurd, given that the latent variable range in most Item Response Theory software is between -6 and 6. Under this uninformative prior, the researchers would also implicitly assume their proposed treatment could cause a *decline* of at least 2 standard deviations in latent ability with prior probability $P(\beta_{TRT} \leq -2) = 49.2\%$. Is it plausible that almost one of every two similarly designed teaching experiments causes this much harm?

Indeed, treatment effects <-1 or >1 are rare in early childhood education (Anderson et al., 2003; Camilli et al., 2010; Tanner-Smith et al., 2018), and so the prior distribution for β_{TRT} should reflect this. For example, a Normal(0, 1.5) prior still conveys that we assume no treatment effect (on average) prior to seeing data, but the prior probability of an effect between -1 and 1 is now $P(-1 \leq \beta_{TRT} \leq 1) = 49.5\%$. In other words, before collecting data, we assume there is an approximately equal chance of a "typical" treatment effect (between -1 and 1) and an "atypical" treatment effect (smaller than -1 or larger than 1). This type of prior is typically deemed

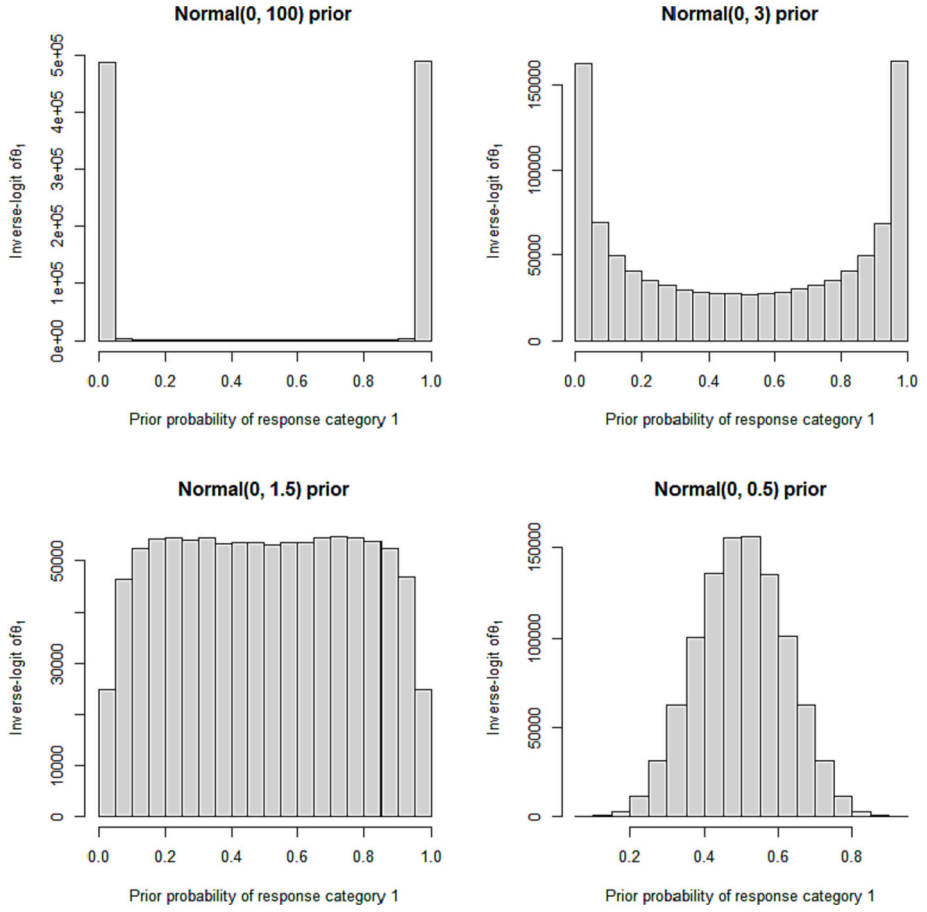


Fig. 2. Illustration of four priors for the first threshold parameter under an inverse-logit transformation in a cumulative logit model.

"weakly-informative", wherein we allow prior knowledge (for example, from a meta-analysis) and/or experience to influence our specification, and then prior uncertainty is deliberately increased to help ensure a conservative result.

Similar logic applies to setting priors for threshold parameters θ_1, θ_2, and θ_3: their plausible prior ranges should reflect common knowledge about the latent variable being modeled, and so there is likely no need for truly *un*informative priors. Under a cumulative logit model, threshold parameters help determine the predicted probability of observing each response category. For example, in the absence of covariates, or with covariates set at their reference values, the predicted probability of observing the lowest ranked category is $\pi_1 = P(Y \leq 1) - P(Y \leq -\infty) = logit^{-1}(\theta_1) - 0 = logit^{-1}(\theta_1)$. This equation highlights the need to pay attention to the transformations within the selected model; here, the *inverse-logit* transformation.

A prior that is uninformative on the latent scale can impose some potentially undesirable features on the probability scale. For example, assigning an uninformative Normal(0, 100) or even a Normal(0, 3) to θ_1 and then applying the inverse-logit transformation results in a U-shaped probability distribution (top left and top right of Fig. 2). Under this prior, extremal probabilities (near 0 and near 1) are approximately equally likely, but almost all other probabilities (>0.1 and <0.9) are nearly impossible—hardly an uninformative distribution. Conversely, assigning a narrow Normal(0, 0.5) prior results in a unimodal, symmetric distribution that essentially rules out probabilities near 0 and 1 *apriori* to data collection (bottom right of Fig. 2). Visualizing the prior shows that if we strive for a weakly-informative prior on the probability scale, we can use a Normal(0, 1.5), which is almost flat between a probability of 0.1 and 0.9, but still allows probabilities near 0 and 1 (bottom left of Fig. 2).

2.4.2.3. MCMC sampling parameters. Our sampling took place using the No-U-Turn Hamiltonian Monte Carlo (NUTS HMC), implemented via the brms R package (Bürkner, 2017), with computational details provided in Appendix B. We used three parallel initially-overdispersed MCMC chains, which were run for 3000 iterations after a 1000-iteration warm-up period with the adaptation parameter adapt "delta" set to 0.85. To our knowledge, there is little consensus around the optimal target acceptance rates (reflected by delta), as these tend to be specific to each estimation task. Our specifications produced satisfactory model convergence with Rhat (Gelman & Rubin, 1992) values of 1.00 for all parameters and Effective Sample Size (ESS) of >500 (and frequently >1000) for all parameters. No divergent transitions occurred for any of the final models presented in the Results section; however, when divergences did take place during model-comparison, there were fewer than 5 and they occurred in the middle of the parameter space and were thus ignored. The Pareto-k diagnostic values (Vehtari et al., 2017) were <0.7 for all models, and thus considered to be satisfactory.

2.4.2.4. Priors and posterior for present study. As is typical, we assume that all prior distributions are *apriori* independent and so the posterior distribution to be sampled is as follows:

$$p(\vartheta|\mathbf{Y}) \propto \sum_{i=1}\sum_{j=1}\sum_{k=1}\sum_{c=1} I\left(Y_{ijk} = c\right) log\left(\pi_{ijkc}\right) \times$$

$$p(\boldsymbol{\Theta})p(\boldsymbol{\beta})p(\boldsymbol{\sigma}_u)p(\boldsymbol{\sigma}_v)p(\boldsymbol{\sigma}_w)p(\boldsymbol{\Omega}_u)p(\boldsymbol{\Omega}_v)p(\boldsymbol{\Omega}_w)$$

We favor weakly-informative priors to complete the posterior. Relative to traditional uninformative priors (e.g., Normal$(0,10^3)$), weakly-informative priors are used to avoid pathological geometries in the posterior that may lead to sampling difficulties. Furthermore, poorly construed uninformative priors can distort estimates more so than reasonable informative priors (McElreath & Koster, 2014). To this end, we assign $N(0,2)$ priors to the threshold parameters $\boldsymbol{\Theta}$ and $N(0,1.5)$ priors to β, both of which are nearly flat on the inverse-logit scale. We assign Half-$N(0,2)$ priors for random effect standard deviation parameters, which are flat near the origin on the logit scale. Finally, we assign uninformative LKJ(1) priors to the correlation matrices, which allow all 3×3 correlation matrices to be equally-likely *apriori* and have the identity matrix as their prior mean.

2.4.2.5. Model averaging in Bayesian analysis. Although it is common to select the "best" or the "preferred" model based on an Information Criterion, less attention is given to the inherent uncertainty in the process of selecting this single model (e.g., Burnham & Anderson, 2004). Competing models may propose equally plausible data-generating processes, and fit the observed data equally well, but use different sets of explanatory variables, and should therefore be considered as formal alternatives. For problem-solving strategies, for example, we do not know whether an adjacent category or a sequential model is the "true" model for the data; yet, it is likely that the estimated treatment effects will be different. It is unclear, therefore, which treatment effect we should take as evidence of experimental efficacy. Indeed, inference on a single model has been deemed "risky" (Hoeting et al., 1999), whereas others note that the process of model selection should at least somewhat diminish effect sizes (Draper et al., 1987; Hodges, 1987).

Bayesian Model Averaging (BMA) is a mechanism that allows pre-specified model parameters to be averaged across a set of competing models, where each model has its own *weight*. In doing so, we employ multi-model inference, incorporating evidence from multiple statistical models to arrive at a result. We apply BMA to the treatment effect to investigate whether a model-averaged treatment effect suggests a different conclusion, compared to the treatment effect from the "preferred" model. Model *weights* can be formulated in different ways, with greater weight given to models that are more probable for the observed data. Most often, this translates into greater weights for models with smaller information criteria among the set of models considered. Here, we follow recent recommendations in Yao et al. (2018) and base our model weights on "stacking weights". Readers are directed to Appendix B for further details.

3. Results

3.1. Model selection

To facilitate model selection, we used the LOOIC, as shown in Table 4. Unlike the often-used Deviance Information Criterion (DIC; Spiegelhalter et al., 2014), LOOIC is fully Bayesian and is more robust than both DIC and WAIC in presence of possible influential observations, weak priors, and finite sample size. Among models with few or no category-specific effects (M1, M2), the cumulative logit model (LOOIC = 9178.5; 270.8 effective parameters) was competitive, reflecting its parsimony relative to the other models. The addition of category-specific treatment effects, as well as class, item, and student category-specific random intercepts, serve to substantially improve LOOIC (M3 through M8), except for a category-specific treatment × pre-sophistication interaction (M5). Formal comparison of M4 vs. M5 yields a LOOIC difference of <0.5SE, and thus we omit the category-specific interaction in subsequent models (M6 through M8). For all except model M8, the sequential model is preferred over the adjacent category model; the latter offers <1SE improvement in LOOIC over the former for model M8. Because the sequential model has a more plausible data-generating mechanism, and the difference in information criteria is not definitive, we select the sequential model over the adjacent category model.

3.2. Conditional treatment effects

Category-specific treatment effects were required to adequately fit our dataset, which indicates that the impact of the treatment was different for each strategy response category. The differential impact of the treatment is clear in the plot of estimated conditional probabilities of using a given sophistication category, adjusted for sophistication at pre-assessment (Fig. 3 and Fig. 4).

Fig. 3 depicts the chances of students employing any given strategy by pre-assessment sophistication. Students who received LT-aligned instruction were less likely to engage in guessing behaviors across all four pre-sophistication levels. For pre-sophistication of L2 or greater, students' modal post-sophistication remained the same. The impact of LT-aligned instruction on other strategies is less clear: students in the LT condition were estimated to use more sophisticated strategies regardless of pre-sophistication. Fig. 4 shows the estimated probability distribution (posterior) of plausible effect sizes. This figure (Fig. 4) highlights two major features, including the (a) different treatment sizes for the three transitions between levels of sophistication and (b) differences between conditional and item-averaged treatment effects.

Conditional on the random effects, the strongest positive treatment effect, and also the only one with a credible interval that excludes 0, occurred for $\beta_{INT\,1}$, which indicates greater chances of transitioning from $Y = 1$ to $Y \geq 2$ in the treatment group (Table 5; top left of Fig. 4). The transition from $Y = 2$ to $Y \geq 3$, given $Y \geq 2$ ($\beta_{INT\,2}$) was unaffected, whereas the transition from $Y = 3$ to $Y = 4$, given $Y \geq 3$ was more likely in the treatment group ($\beta_{INT\,3}$), but its credible interval contains 0. In Bayesian settings, we obtain a distribution of the associated treatment effects; using these distributions, we determine the posterior probabilities of positive treatment effects to be: $P(\beta_{INT\,1} > 0) = 100\%$, $P(\beta_{INT\,2} > 0) = 65.3\%$, and $P(\beta_{INT\,3} > 0) = 86.7\%$.

3.2.1. Model-averaged conditional treatment effects

The magnitude of treatment effects varied depending on which of item, student, and classroom category-specific random intercepts were included. To provide an estimate of the treatment effect that is robust to model selection, we performed model averaging of all models with category-specific random effects (M6, M7, M8 in Table 4). In decreasing order, the optimal non-zero model weights were 0.546, 0.409, and 0.044 for adjacent-category M8, sequential M8, and sequential M6, respectively. Based on 5000 MCMC samples, model-averaged probabilities of a positive treatment effect, conditional on the random effects, were 100%, 65.6%, and 86.8%, for $\beta_{INT\,1}$, $\beta_{INT\,2}$, and $\beta_{INT\,3}$, respectively.

Table 4
Leave-one-out Information Criteria (LOOIC) with number of effective parameters (in parentheses).

Model	Category-specific Effects	Sequential	Adjacent Category
M1	–	9261.6 (272.2)	9305.1 (265.0)
M2	Interv.	9197.0 (266.2)	9231.9 (258.6)
M3	Pre-soph.	8659.5 (270.2)	8695.8 (263.0)
M4	Interv. + Pre-soph.	8580.0 (266.7)	8607.6 (260.1)
M5	Interv. × Pre-soph.	8576.3 (275.9)	8605.2 (267.3)
M6	Interv. + Pre-soph. + Class	8519.7 (289.3)	8545.1 (277.0)
M7	Interv. + Pre-soph. + Class+Item	7588.0 (322.2)	7653.3 (315.6)
M8	Interv. + Pre-soph. + Class+Item+Student	7018.9 (526.3)	7002.9 (519.1)

Note. Smaller LOOIC values indicate the preferred model; fewer effective parameters indicate lower model complexity. Interv. denotes the binary treatment variable; Pre-soph. Denotes each student's 4-level pre-assessment strategy sophistication; Class, Item, and Student denote category-specific random intercepts. The Cumulative model (LOOIC = 9178.5; 270.8 effective parameters) contains pre-sophistication, intervention, class, item, and student effects, but does not allow any category-specific effects, so it was not listed in the table.

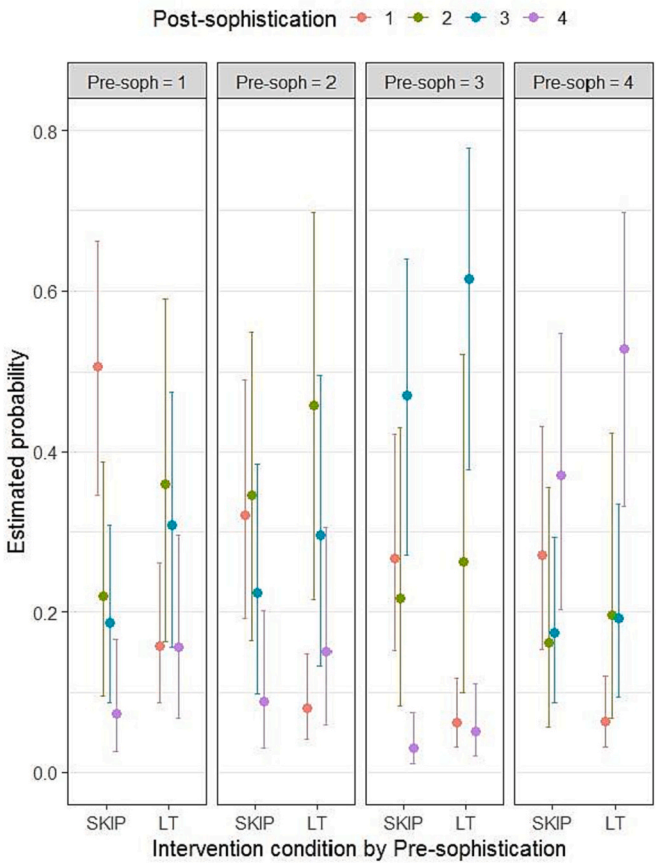


Fig. 3. Estimated conditional probabilities of post-assessment strategy sophistication by treatment condition (SKIP versus LT) and pre-assessment sophistication.
Note. Points represent posterior medians with 95% Credible Intervals.

3.3. Population-averaged treatment effects

By definition, conditional treatment effects apply to the "typical" item, student, and classroom. Item effects are commonly interpreted in terms of difficulty, which is intuitive when one is primarily concerned with correct responses. It may be possible to extend this interpretation onto item sophistication, but we feel it is more accurate to amend it to "items requiring typical levels of sophistication". Regardless, we may wish to broaden our scope of inference to the population from which our 20 items came from. Formally, conditioning the treatment effects on the typical item requires that we hold item characteristics fixed, for example: number range, position of the unknown quantity, verbal complexity, etc. With this in mind, it is not clear what constitutes a "typical" item and motivates the estimation of an item-averaged treatment effects.

Population-averaged (PA) estimates are attenuated relative to conditional estimates, where the degree of attenuation depends on random effect variance (e.g., Zeger et al., 1988). Averaged over items, but for the typical student in a typical classroom, the PA treatment effects were: $\beta^{PA}_{INT1} = 1.43(0.09, 3.66)$, $\beta^{PA}_{INT2} = -0.81(-2.87, 4.94)$, and $\beta^{PA}_{INT3} = 0.36(-1.66, 1.84)$. Accordingly, the posterior probabilities of a positive PA treatment effect were 98.3%, 37.1%, and 63.6%, respectively.

We compare the posterior distributions of conditional and item-averaged treatment effects in Fig. 4 (left versus right columns). The attenuation of treatment effects towards 0 is apparent in the item-averaged posteriors, as is the apparent bi-modality of the treatment effect. For example, for the transition from $Y = 2$ to $Y \geq 3$, there were two apparent sub-populations: (a) students who transitioned to more sophisticated strategies post-treatment, and (b) students who remained at $Y = 2$ post-treatment. Correspondingly, the estimated between-item standard deviation for this transition (σ_{u2}) of 2.48(1.79, 3.42) (Table 5) was the largest among all random effect standard deviations.

3.4. Student and item effects

We found evidence that there exist all three of class, student, and item category-specific random intercepts in our data. It follows that we can compute several category-specific Intra-Class Correlation (ICC) coefficients, which quantify the similarity of sophistication

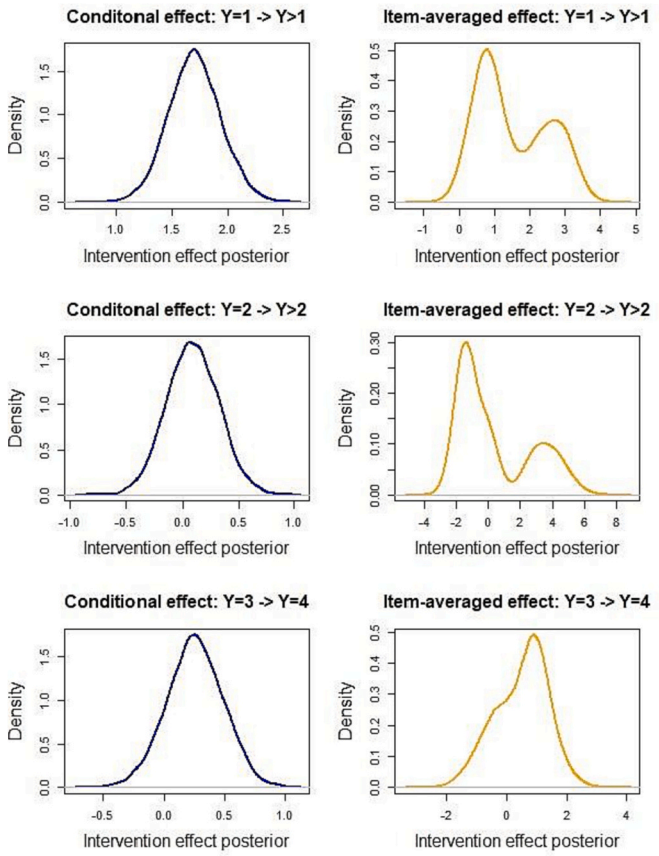


Fig. 4. Posterior distributions of conditional treatment effects (left) and item-averaged treatment effects (right) for each of three latent transitions under the sequential logit model (M8).
Note. Treatment effects are shown on the logit scale, where 0 indicates no effect.

ratings within each level of hierarchy and response category. For example, at the classroom level the ICC is $\frac{\sigma_w^2}{\sigma_w^2+\sigma_v^2+\sigma_u^2+\pi^2/3}$, at the student level the ICC is $\frac{\sigma_v^2}{\sigma_w^2+\sigma_v^2+\sigma_u^2+\pi^2/3}$, and at the level of students within a given classroom the ICC is $\frac{\sigma_w^2+\sigma_v^2}{\sigma_w^2+\sigma_v^2+\sigma_u^2+\pi^2/3}$ (Raman & Hedeker, 2005; Snijders & Bosker, 2011).

At the classroom level, the largest ICC of 0.14(0.06, 0.30) was recorded for the transition from $Y = 3$ to $Y = 4$, among those who reached at least $Y = 3$. The remaining two ICCs were minimal with posterior medians of < 0.05. The three classroom random effects were positively correlated (Table 5), indicating they likely measure the same latent classroom-level variable. We interpret this variable to be a measure of socio-demographic characteristics because the 16 classrooms comprised four school districts that drew enrollment from economically, racially, and ethnically diverse areas of the city.

At the student level, regardless of classroom, ICCs were more substantial with estimates of 0.39(0.32, 0.46), 0.19(0.12, 0.27), and 0.25(0.17, 0.34) for the three latent transitions, respectively. Unlike classroom random effects, only the effects that govern latent transitions at the higher range of sophistication were positively correlated. This correlation may reflect general motivation that naturally varies by student, or other factors, such as the quality of the student-teacher relationship.

We modify the usual interpretation of student random effects from "latent ability" to "latent sophistication" because it is not clear that students with greater "ability" - determined by correctness - necessarily choose high-level sophistication strategies. However, we show that latent sophistication was positively correlated with correctness (Fig. 5), where the latter was obtained using a classic Rasch score, as described in Clements et al. (2020). In the case of a single student random intercept across all three latent transitions (model M7), latent sophistication had a moderate positive correlation with correctness-based Rasch factor scores for arithmetic ($R = 0.58$ (0.50,0.65)). When we allowed student random intercepts to vary by latent transition via a category-specific effect (model M8), latent sophistication was less correlated with correctness, with diminishing correlation as transitions occurred at higher-sophistication response categories. This suggests that latent sophistication measures a distinct construct relative to the Rasch score when transitions occur between higher levels of sophistication.

Item-level ICCs were similar to student-level ICCs, except for the transition from $Y = 2$ to $Y \geq 3$, which had the largest estimate of 0.49(0.34, 0.65). This estimate is also reflected in the large random effect standard deviation estimate and the apparent bi-modality in the item-averaged treatment effects (Fig. 4). Item random effects were correlated such that items with large thresholds for the $Y = 3$ to

Table 5
Conditional posterior medians (95% Credible Intervals) on the logit scale of threshold parameters and treatment effects.

		Sequential Model (M8) Posterior Estimate (CrI)	Adj. Category Model (M8) Posterior Estimate (CrI)
Thresholds	θ_1	0.03 (−0.64, 0.67)	1.34 (0.53, 2.16)
	θ_2	−0.19 (−1.34, 0.94)	0.35 (−0.66, 1.38)
	θ_3	0.93 (0.16, 1.70)	0.84 (0.14, 1.55)
Intervention	β_{INT1}	1.70 (1.24, 2.18)	1.78 (1.26, 2.29)
	β_{INT2}	0.09 (−0.36, 0.56)	−0.22 (−0.57, 0.13)
	β_{INT3}	0.26 (−0.20, 0.71)	0.16 (−0.27, 0.58)
Pre-soph. = 2	β_{PRE21}	0.77 (0.47, 1.07)	0.77 (0.43, 1.11)
	β_{PRE22}	−0.27 (−0.62, 0.07)	−0.18 (−0.55, 0.18)
	β_{PRE23}	0.00 (−0.51, 0.53)	0.13 (−0.34, 0.57)
Pre-soph. = 3	β_{PRE31}	1.04 (0.71, 1.36)	0.49 (0.06, 0.93)
	β_{PRE32}	0.66 (0.27, 1.08)	1.13 (0.74, 1.50)
	β_{PRE33}	−1.80 (−2.28, −1.35)	−1.65 (−2.05, −1.23)
Pre-soph. = 4	β_{PRE41}	1.02 (0.66, 1.38)	0.23 (−0.22, 0.69)
	β_{PRE42}	1.05 (0.64, 1.48)	0.13 (−0.033, 0.57)
	β_{PRE43}	1.68 (1.23, 2.13)	1.67 (1.26, 2.07)
Classrooms	$\sigma_{\omega 1}$	0.52 (0.22, 0.92)	0.31 (0.02, 0.73)
	$\sigma_{\omega 2}$	0.75 (0.41, 1.20)	0.35 (0.06, 0.67)
	$\sigma_{\omega 3}$	1.01 (0.60, 1.58)	0.91 (0.53, 1.47)
	$\rho_{\omega 12}$	0.69 (0.08, 0.97)	0.21 (−0.69, 0.91)
	$\rho_{\omega 13}$	0.64 (0.03, 0.96)	0.43 (−0.49, 0.95)
	$\rho_{\omega 23}$	0.68 (0.18, 0.95)	0.36 (−0.41, 0.90)
Students	$\sigma_{\upsilon 1}$	1.77 (1.56, 2.00)	1.94 (1.70, 2.19)
	$\sigma_{\upsilon 2}$	1.55 (1.32, 1.79)	1.02 (0.84, 1.22)
	$\sigma_{\upsilon 3}$	1.32 (1.05, 1.61)	1.28 (1.04, 1.54)
	$\rho_{\upsilon 12}$	0.13, (−0.05, 0.30)	−0.38 (−0.56, −0.18)
	$\rho_{\upsilon 13}$	0.11 (−0.10, 0.32)	−0.22 (−0.43, −0.01)
	$\rho_{\upsilon 23}$	0.68 (0.50, 0.83)	0.36 (0.09, 0.61)
Items	σ_{u1}	1.10 (0.78, 1.55)	1.70 (1.25, 2.32)
	σ_{u2}	2.48 (1.79, 3.42)	2.22 (1.65, 2.99)
	σ_{u3}	0.94 (0.62, 1.37)	0.84 (0.57, 1.23)
	ρ_{u12}	−0.01 (−0.42, 0.40)	−0.77 (−0.91, −0.52)
	ρ_{u13}	0.52 (0.09, 0.81)	0.51 (0.12, 0.80)
	ρ_{u23}	−0.34 (−0.69, 0.10)	−0.44 (−0.75, −0.02)

Note. The estimates were adjusted for pre-assessment strategy sophistication, with random effect standard deviations and between-response-category correlations.

$Y = 4$ transition tended to also have large thresholds for the $Y = 1$ to $Y \geq 2$ transition, but small thresholds for the $Y = 2$ to $Y \geq 3$ transition.

To investigate the item effects, we plot the estimated probabilities of using different levels of sophistication by item and treatment condition, accounting for pre-assessment sophistication (Fig. 6). In this figure, items are shown in column panels that are ordered according to an empirically validated developmental progression, such that Item 1 is least difficult and Item 72 is most difficult. For easy items, children in the LT condition consolidate their strategies into higher-level categories relative to their SKIP condition peers. For more difficult items, children in the LT condition diversify their strategies, whereas children in the SKIP condition consolidate their strategies to $Y = 1$. This difference in strategy breadth for difficult items is likely responsible for the bi-modality in the item-averaged treatment effects (Fig. 4; right column).

4. Discussion

Accuracy and fluency are important to track and are usually straightforward to analyze – there is clear guidance for how to implement and interpret the necessary statistical models. We advocate that problem-solving strategy sophistication is equally informative and important to track in the work of teaching and school psychology. However, due to its ordinal, categorical nature, strategy sophistication requires more nuanced statistical analysis plans. As a first step, this study offers a case where (a) learning trajectories (LTs) served as a pedagogical tool that orders the relative sophistication of children's observed problem-solving behaviors, as well as (b) describes how to implement and interpret statistical models for this outcome.

To illustrate the utility of strategy sophistication, we leverage data from an experiment that examined the contributions of LTs and LT-aligned instruction to arithmetic learning outcomes. LTs are a curriculum-agnostic pedagogical tool that was used to improve arithmetic accuracy, as well as increase the modal sophistication of kindergarten students' problem-solving strategies across a variety of increasingly complex story problems (i.e., applied problems). In the course of the analysis, we show that common ordinal models do not accurately fit our strategy dataset, such that they fail to attain sufficient specificity in describing more nuanced, but important effects of LT-aligned instruction (referred to as the treatment effect).

Our modeling approach enabled us to investigate the treatment effect in four novel ways: (a) using probabilistic statements about efficacy, we report the probability that LT-aligned instruction increased students' strategy sophistication; (b) quantifying this effect for

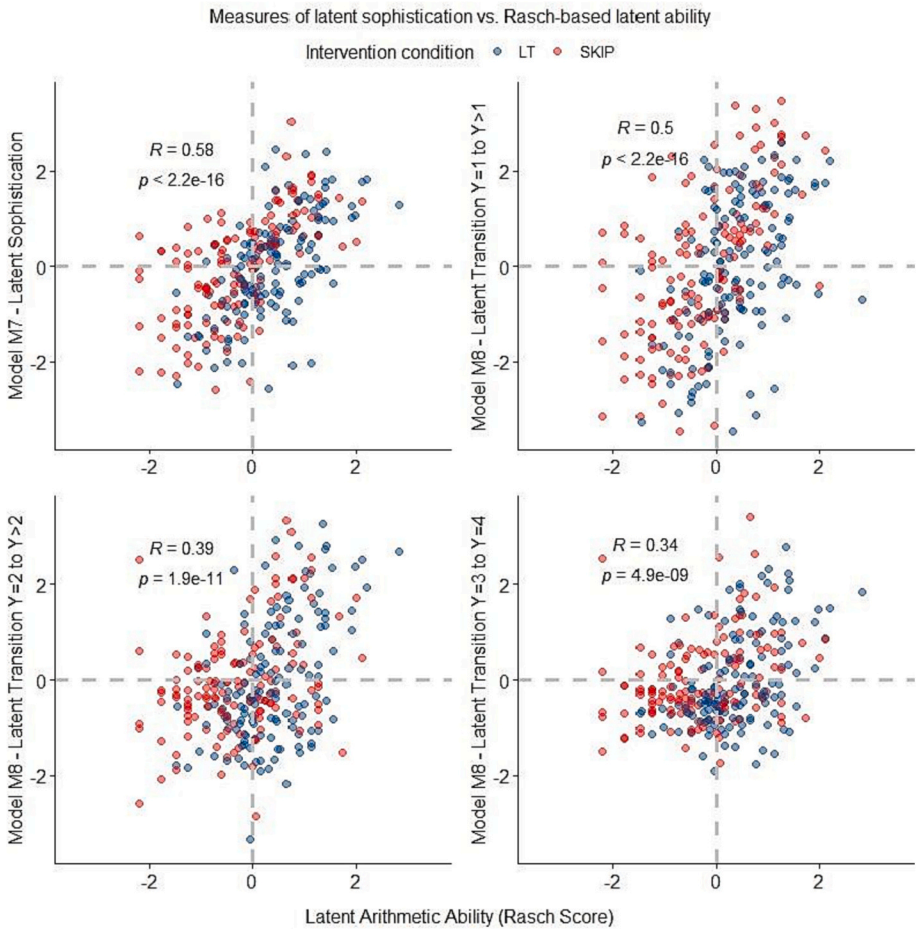


Fig. 5. Correlation between Rasch-based latent arithmetic factor scores with measures of latent sophistication from sequential models M7 (top left) and M8 (top right, bottom left, bottom right).
Note. Model M7 features a single student random intercept across all three latent transitions; model M8 has three correlated student random intercepts - one per transition. Each point is a unique student and is colored based on treatment condition.

the typical item, student, and classroom (conditional effect), as well as averaged over the population of items (marginal effect); (c) showing that the treatment effect varies for each transition between successive strategy sophistication ratings; and (d) investigating the sensitivity of the treatment effect to a particular statistical model and report both model-averaged and model-dependent estimates. We further detail the contributions of our work in a methodological context, treatment-specific context, and in the context of interpreting latent student and item effects. Additionally, we offer implications for the use of problem-solving strategies as an early math outcome in relation to the work of school psychologists.

4.1. Methodological implications

The choice of statistical model – here, the type of multilevel ordinal model – determines not only the software needed, but also posits a process by which strategy sophistication data came to be. This process is called the "data-generating process" in statistical literature. More flexible models often fit better, but we urge the reader to pay closer attention to what each model assumes about the nature of children's problem-solving behaviors.

For example, both the adjacent category and sequential models allow category-specific effects, both fit better than the model without category-specific effects, and all three can be implemented using the same R package. Yet, it is the sequential model that produces the most interpretable data-generating process. Permit us a metaphor: imagine a racetrack with hurdles that a runner must clear. A model without category-specific effects assumes the hurdles are arranged according to increasing height and each runner has some (latent) ability to progress along the track. However, this model does not prescribe a direction to run in, although we assume is it better to be further along the track. Conversely, the sequential model with category-specific effects does not assume that the hurdles are ordered by height, but it prescribes a single direction in which runners proceed. Unlike the prior model, the sequential model assumes that we must consider the runner's unique (latent) ability to clear each hurdle. Combined with a prescribed direction of

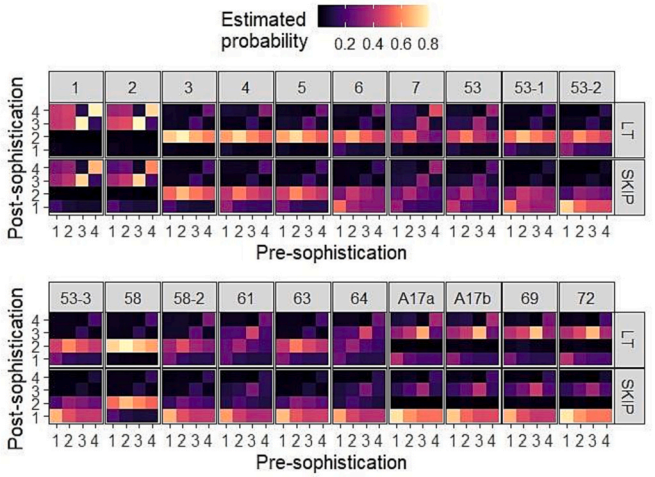


Fig. 6. Conditional posterior median probabilities of strategy sophistication at pre- and post-assessment, by treatment condition (row panels) and Item (column panels).

Note. Items are arranged according to their developmental progression from easiest (Item 1) to most difficult (Item 72). Within each panel, the position of the colored square indicates pre-post growth in strategy sophistication. For example, for Item 1, students with pre-sophistication of L1 or L2 were most likely to use strategies L3 or L4 at post-assessment in both experimental conditions. In contrast, for Item 61, students in the LT condition were likely to remain at their pre-sophistication levels, whereas students in the SKIP condition were most likely to revert to L1 strategies at post-assessment.

motion, if we find a runner further down the track, we may infer that they have successfully cleared each preceding hurdle to get there.

Indeed, no data-generating process is infallible; however, we consider the sequential model to be more tenable when we refer to the robust research that forms the basis for learning trajectories (for a comprehensive review, see Clements & Sarama, 2021, as well as Sarama & Clements, 2009). Taken together, the sequential model fits the data better and is grounded in mathematics education research, giving us confidence in this data-generating process.

4.2. *The necessity of category-specific effects to capture the nature of early arithmetic strategies*

In the present study, the sophistication scale is composed of four categories, which enabled us to estimate category-specific effects using relatively few parameters. More parsimonious cumulative logit models may become more viable if the sophistication scale grows to include more categories. That is, with 10 or more categories of sophistication, models with category-specific effects would require many more parameters and may be over-parameterized, wherein the cumulative model may be superior to the sequential model.

The need for category-specific effects is strongly supported by our understanding of the skills required to solve early arithmetic problems. Each level of strategy sophistication (L1–L4) is linked to observable arithmetic problem-solving behaviors that are built on interrelated, but distinct numeracy principles. For example, L2 indicates the "count all" strategy was observed: students demonstrated working knowledge of counting principles, such as one-to-one correspondence, stable-order, and order-irrelevance (Gelman & Gallistel, 1986; Gelman & Meck, 1983). L3 is built on the competences of the previous levels and indicates the "count on" strategy was observed: students demonstrated working understanding of cardinality and conservation (Fuson, 1992, 2012). L4 strategies are an outgrowth of interconnected knowledge about numbers: students demonstrated more complex counting strategies, known combinations, as well base-10 knowledge for decomposing (Baroody, 2006). Category-specific effects are needed to adequately model the dynamic knowledge and skills that serve as the catalyst for each of these transitions.

The hierarchical nature of our data brings into focus the contrast between conditional and population-averaged (marginal) treatment effects. When we take the scope of inference to be the population of items, we find evidence of apparent bi-modality in the item-averaged marginal treatment effects, such that some students make the transition to higher-level sophistication post-treatment, whereas others do not. In other words, we were able to identify that this apparent bifurcation is most pronounced at the transition between $Y = 2$ and $Y \geq 3$: from "count all" to "count on", and from "count on" to "known combinations" or "decomposing number." Further work is needed to determine which item characteristics encourage transitions to higher levels of sophistication.

Different statistical models imply different directions, magnitudes, and scope of the treatment effect. For example, whether the treatment effect (LT-aligned instruction) is strategy-specific and whether the treatment effect is conditional or marginal depends on which of our eight models is used. Like many other researchers do, we select just one model to describe the treatment effect – the best-fitting model that also features the lowest information criterion. However, in doing so, we implicitly disregard the uncertainty around selecting this final model, as well as any conflicting evidence brought forth by other models fit to the data. One standard method of incorporating all available evidence regarding treatment is by using Bayesian model averaging, as demonstrated in this article. In employing model averaging, we incorporate information about LT-aligned instruction from all models with category-specific effects without relying on any one as the "true" model. In doing so, we disconnect the treatment effect from the assumptions of any single

model, thereby delivering a more robust metric.

4.3. Implications for student effects: Correlation between strategy sophistication and Rasch scores

The interpretation of student effects in the context of strategy sophistication is nuanced and depends in part on whether these are category-specific (as in our model M8) or averaged across categories (as in our model M7). We call these student effects "latent sophistication", and in the latter case, show that these effects share a moderate amount of information with traditional accuracy-based Rasch scores. This suggests there is utility in using latent sophistication in conjunction with Rasch scores and that these are complementary measures of mathematical competences. Further concurrent and predictive validation exercises will help delineate areas of similarity and difference between these two measures. For instance, the correlation between latent sophistication and Rasch scores differed somewhat by classroom, suggesting that different student characteristics and learning environments may moderate this relationship. In the case of category-specific student effects, the correlation with Rasch scores is weaker and more difficult to interpret because we obtain three sets, each governing a different latent transition. More research is required to determine what can be gleaned from studying latent transitions between strategy sophistication categories and how these relate to traditional Rasch scores. In fact, if item responses are products of thinking, the measurement of problem-solving processes can potentially link errors to strategies that are lower in sophistication.

Students in our experiment were unable to use instructor feedback or previous trial-and-error to inform their strategy selection due to the nature of assessment. Future experiments can track how strategy preferences evolve over time - between and within instructional sessions - and how prior success with a given set of strategies informs future strategy preferences. It is also plausible there are different groups of students, including those who are mathematically "adventurous" and will attempt to use a variety of strategies and be less discouraged by mistakes (Kerkman & Siegler, 1993). There may also be those who are mathematically "conservative" and will rely on strategies that maximize their chances of obtaining the correct answer (Kerkman & Siegler, 1993). Latent profile analyses can be used to investigate the prevalence of such groupings.

4.4. Implications for item effects: Changing strategies in response to the demands of story problems

The first five items are categorized as *Result Unknown* story problems, where the location of the unknown quantity is defined by the total in the set (e.g., $2 + 1 = x$). For these items, pre-post growth in strategy sophistication is similar between experimental conditions. Item 6 through Item 58–2 are categorized as *Find Difference* story problems, where the location of the unknown quantity is defined by the addend or subtrahend (e.g., $5 + x = 7$). Except for Item 58, students in the LT condition relied less on guessing behaviors relative to their SKIP peers. For the LT condition, if pre-sophistication was L3, a bifurcation can be observed, such that the student remained at L3 or transitioned into L4 by post-assessment. Interestingly, for students with pre-sophistication at L4, there was equal chance of remaining at L4 or curtailing to L2 at post-assessment. This growth pattern is reflected in the bi-modal item-averaged treatment effects shown in Fig. 4 (right column).

Taken together, we show evidence that students are adapting their strategies in response to the demands of the story problems. This is evident in the patterns of change (within experimental condition) between the problem types (i.e., Result Unknown, Find Difference, and Start Unknown). For example, for Result Unknown problems, LT students are shifting to more sophisticated strategies at post-assessment. For the more difficult Find Difference problems, LT students are using L2 (count all), regardless of pre-sophistication. For the most difficult Start Unknown story problems, or items that explicitly assess more sophisticated counting on or skip counting knowledge (e.g., "What is 49 almonds and 10 more?"), LT students using L1 strategies are likely to remain at L1 or jump to L3. Students using L2 or higher at pre-assessment are likely to use L3 strategies. In contrast, students in the SKIP condition do not adapt as much as their LT peers and primarily rely on guessing behaviors (L1) or count all (L2) for all but the easiest items.

Our findings suggest that the type of story problem (i.e., Result Unknown, Find Difference, and Start Unknown) contributes to modal strategy sophistication *and* range of strategies deployed. A future study might explore ways of quantifying the effect of story problem type, strategy variability (i.e., range), and its relationship to strategy sophistication, as well as overall arithmetic competence.

4.5. Implications for practice: Learning trajectory-aligned instruction and activities strengthen arithmetic strategy sophistication

We compared three families of ordinal logistic models according to their theoretical utility and goodness-of-fit for the observed data. We found that the sequential logit model was selected using information criteria, which suggests that children ascend through increasingly sophisticated strategies in a forward, sequential order. This is broadly consistent with the theoretical foundations of the developmental progressions (Clements & Sarama, 2021; Sarama & Clements, 2009). Of course, one should not over-interpret the sequential model as a metaphor for all learning. Children's knowledge and skills can be fragile for multiple reasons and a "descent" down the sophistication scale (i.e., strategy curtailment) routinely occurs (Geeslin & Graham, 1992; Kieren & Pirie, 1992). However, the sequential logit model cannot accommodate this type of reversal (Peyhardi et al., 2015) and so these transitions must be modeled in other ways, a limitation of the selected model. That said, any given student is predicted to use all four levels of sophistication with some probability at pre- and post-assessment within our analysis. To allow the aforementioned reversal to formally occur, one could explore the use of multi-state models (e.g., Hougaard, 1999), which falls outside the scope of this article.

An anonymous reviewer suggested one possible reason why the same or lower level of sophistication may be observed at post-assessment: LT-aligned instruction reduces fluctuation between levels of sophistication, allowing students to consolidate concepts that define particular arithmetic strategies. As an example, consider the transition from L3 (count on) to L4 (known combinations, skip

counting strategies, and decomposition). Students operating at L3 at pre-assessment had opportunity to consolidate their count on (concrete and abstract, forwards and backwards) skills as they acquired the cardinality principle (Fuson, 1992, 2012; Sarnecka & Carey, 2008). Yet, their knowledge of combinations and decomposition skills may still be fragile; thus, they fall back on strategies that are lower in sophistication (but still reflect complex understanding of numeracy concepts and principles). This interpretation is also consistent with strategy development theories that suggest that multiple strategies exist and compete with each other; the most sophisticated skills do not always dominate, although they may be present (e.g., Siegler, 1995, 2007).

It is instructive to interpret the significance of our findings with an eye towards Clements et al. (2020), who focused exclusively on correctness and report an effect size of 1.08 in favor of the students who received LT-aligned instruction (versus SKIP instruction). Like many randomized teaching experiments describing efficacy, the difference between experimental conditions in arithmetic learning was based on correctness expressed as a Rasch score. Our findings in the current article provide a more granular view into the efficacy of an LT-aligned approach to instruction, which centers on the arithmetic strategy sophistication across an increasingly complex variety of problems.

Worth noting is the composition of SKIP students in Quadrants 1 (top left) and LT students in Quadrant 4 (bottom right) within each panel of Fig. 5. Within Quadrant 1, SKIP students had lower levels of correctness, but higher levels of sophistication relative to the average. Recall that LT and SKIP conditions both had one-on-one instructional sessions with LT-based activities. However, the SKIP condition was defined by opportunities to solve problems up to $N + 2$ levels above their current levels of thinking (N). Quadrant 1 suggests that encountering more complex arithmetic problems may have elicited more sophisticated strategy behavior, but students were not yet able to unpack the mechanics that would yield the correct answer.

In summary, three lines of evidence suggest that LT-aligned instruction can support the development of more sophisticated arithmetic strategies. First, using trial and error (L1) was reduced post-experiment for both conditions, but was more greatly reduced for students in the LT condition. Second, students using the "count on" strategy (L3) were more likely to transition into using the most sophisticated strategy (L4) by post-assessment, while still maintaining accuracy across increasingly complex problem types (Clements et al., 2021). Third, regardless of whether students transitioned above the "count all" (L2) strategy, all students were more accurate post-treatment across increasingly complex arithmetic problems. Taken together, these findings on correctness and sophistication serve as a multi-dimensional argument in favor of using the LT-aligned approach relative to teach-to-target skills (i.e., the SKIP approach).

4.6. How the findings of this study can support the work of school psychologists

Fluency, or procedural fluency, is a critical component of mathematics proficiency and defined as "the ability to apply procedures accurately, efficiently, and flexibly; to transfer procedures to different problems and contexts; to build or modify procedures from other procedures; and to recognize when one strategy or procedure is more appropriate to apply than another" (NCTM, 2014). Our findings suggest fluency is cultivated by advancing through all levels of strategy sophistication to build conceptual understanding (NCTM, 2000, 2014), which is critical information for early elementary teachers to understand when trying to meet individual student needs.

An understanding of how to situate arithmetic problem-solving behaviors by relative sophistication is therefore actionable information for school psychologists working with teachers and academic interventionists (e.g., math specialists) through channels of consultation and feedback. School psychologists can support teachers as they look for evidence of student thinking embedded within problem-solving strategies, interpret what that behavior signals about their conceptual understanding, and then make decisions about how to differentiate instruction based on that interpretation. Tier 2 interventionists and tutors working to support student understanding of number relations (Fuchs et al., 2005; Fuchs & Vaughn, 2012) may find this work useful for deeper understanding of diagnostic assessment, such as error analysis (Lembke et al., 2012). Indeed, understanding patterns of student error or misconception in coordination with patterns strategy sophistication offers more precise information about why particular errors occur and what concept or skill is a productive target for instruction. Further, triangulating strategy sophistication can yield insight into the specific concept to target for more intensive instruction (Allsopp et al., 2007).

4.7. Limitations and future directions

We describe two types of limitations to this work: those associated with the statistical models and those associated with the implications for practice. The sequential model – selected as the "final" model in this study – does not explicitly allow for a decrease in sophistication. Of course, such regression is routinely expected to occur and indeed happens in our data. Although sequential models best explain the patterns in our case, interested readers may find greater utility in multi-state models (e.g., Jackson, 2011) that can be used to explicitly quantify the probabilities of all possible transitions between response categories. Furthermore, whereas category-specific effects are needed for our data, these models may be overparameterized in cases with many response categories. Readers are urged to compare models with category-specific effects to the traditional cumulative ordinal models in terms of information criteria. Category-specific effects that are similar in magnitude can be effectively "collapsed" into the effects accommodated by the traditional cumulative logit model.

We also caution against replicating this experiment in schools, despite its resemblance to Tier 2 services, such as small group work and tutoring (Fuchs et al., 2012). Although Clements et al. (2020) offered implementation details, a sequence of pedagogical activities developed by researchers in collaboration with practitioners, and positive effect sizes, the original efficacy study was conceived as testing a theoretical assumption underlying the learning trajectories.

Future directions might more intentionally examine what factors contribute to variability in arithmetic strategy sophistication that

are malleable to instruction. A particularly fecund area is learning to model and solve word problems (Verschaffel et al., 2000)—specifically how features of the mathematical and semantic structure, alongside story context (Carpenter, 1985; Carpenter et al., 2015; Fuchs et al., 2020), serve to help or hinder students' transition from less to more sophisticated strategies (Greer, 1997; Lesh & Lamon, 2013). Indeed, teachers and school psychologists are well-positioned to support children as they construct mathematical models and interpret the outcome of computational work. Another potential line of inquiry lies with the contribution of achievement-emotions to mathematics outcomes, including fluency and strategy sophistication. McLeod's (1992) seminal paper on affect (i.e., beliefs, attitudes, and emotions) in mathematics education states that "emotional reactions" to mathematics have not received much attention, standing in contrast to studies reporting the association between self-regulation and early math skills. Investigating the prevalence of positive and negative discrete emotions *during* problem solving *in situ* can expand our understanding of variability in mathematics outcomes.

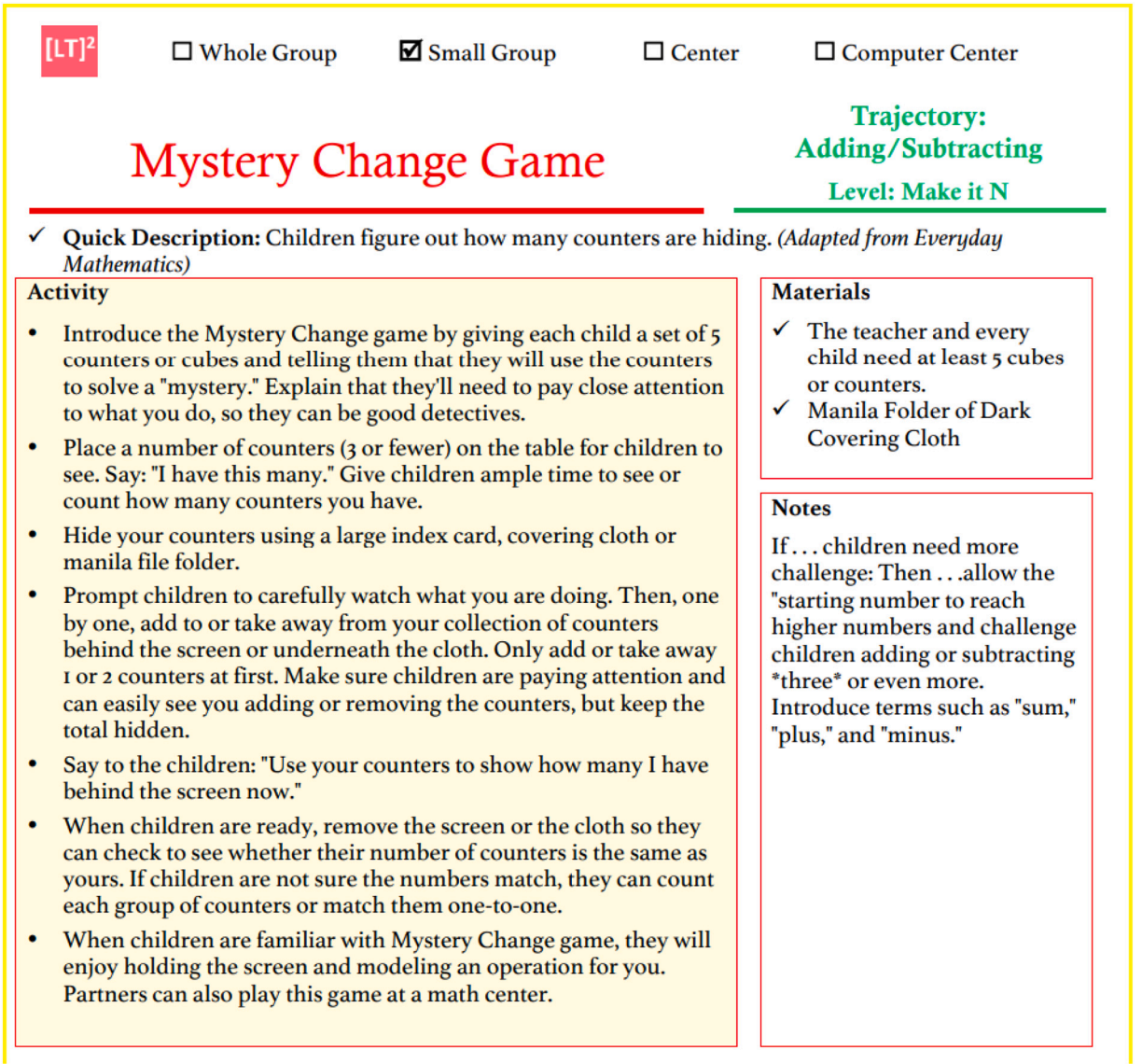
Acknowledgements

The authors thank three anonymous reviewers and the two Editors for their insightful comments and suggestions, as well as Menglong Cong, Kayla McCreadie, Dr. Jason Downer, and Dr. Bonnie Nastasi for their contributions and helpful discussion around the work of school psychologists. The authors were supported in part by IES grants #R305A150243 and #R305A200100.

Appendix A

A.1. Sample instructional activity

The following activity is from the free, publicly available LT^2 website: https://www.learningtrajectories.org/learning_trajectories.



[LT]² ☐ Whole Group ☑ Small Group ☐ Center ☐ Computer Center

Mystery Change Game

Trajectory: Adding/Subtracting

Level: Make it N

✓ **Quick Description:** Children figure out how many counters are hiding. *(Adapted from Everyday Mathematics)*

Activity

- Introduce the Mystery Change game by giving each child a set of 5 counters or cubes and telling them that they will use the counters to solve a "mystery." Explain that they'll need to pay close attention to what you do, so they can be good detectives.
- Place a number of counters (3 or fewer) on the table for children to see. Say: "I have this many." Give children ample time to see or count how many counters you have.
- Hide your counters using a large index card, covering cloth or manila file folder.
- Prompt children to carefully watch what you are doing. Then, one by one, add to or take away from your collection of counters behind the screen or underneath the cloth. Only add or take away 1 or 2 counters at first. Make sure children are paying attention and can easily see you adding or removing the counters, but keep the total hidden.
- Say to the children: "Use your counters to show how many I have behind the screen now."
- When children are ready, remove the screen or the cloth so they can check to see whether their number of counters is the same as yours. If children are not sure the numbers match, they can count each group of counters or match them one-to-one.
- When children are familiar with Mystery Change game, they will enjoy holding the screen and modeling an operation for you. Partners can also play this game at a math center.

Materials

✓ The teacher and every child need at least 5 cubes or counters.
✓ Manila Folder of Dark Covering Cloth

Notes

If . . . children need more challenge: Then . . .allow the "starting number to reach higher numbers and challenge children adding or subtracting *three* or even more. Introduce terms such as "sum," "plus," and "minus."

Although this is listed as a "Small Group" activity, it was implemented within the one-on-one instructional sessions. This particular game was particularly useful for those students preparing to transition from find result problems ($5 + 2 = x$) to find change/difference problems ($5 + x = 7$).

A.2. Sample item

Materials Needed: 10 tokens.

Trial A. Angie bought some candies. Her mother bought her 3 more candies. Now Angie has 5 candies. How many candies did Angie buy?

Trial B. Say, Blanca had some tokens. She lost 2 tokens playing. Now she has 7 tokens. How many tokens did Blanca have before she started to play?

A.3. Scoring

0 = Incorrect.
1 = Correct.
9 = No response.

A.4. Strategies (What did the child do?)

1 = child identifies or makes a set of objects in answer to question, but does not verbalize. response.
2 = counting on and keeping track (e.g., keeping track with fingers or keeping track with. counts).
3 = counting using objects to keep track (e.g., direct modeling).
4 = verbalized derived combination (e.g., "3+3 is 6, and plus 1 is 7").
5 = verbalized combination (e.g., "2+3 is 5").
7 = Other strategy.
8 = Strategy not observed.
9 = NA.

Appendix B

B.1. Computational details

Estimation took place by sampling from the posterior distribution using the Adaptive No-U-Turn variant of Hamiltonian Monte Carlo (NUTS HMC) in Stan (Hoffman & Gelman, 2014) implemented using the brms R package (Bürkner, 2017). HMC produces new proposals by simulating particle motion in a frictionless **q**-dimensional Hamiltonian system, where **q** is the total number of parameters to be sampled. Compared with other samplers, such as Gibbs and Metropolis-Hastings, HMC produces nearly uncorrelated samples, with much greater Effective Sample Size (ESS) per minute (Carpenter et al., 2017; Nishio & Arakawa, 2019), and also tends to be robust to high-dimensional and elliptical posteriors relative to the Gibbs sampler (Monnahan et al., 2017). Furthermore, unlike Gibbs, HMC does not require conjugate priors, which can facilitate more flexible model specification and selection of reasonable hyper-parameters for prior distributions.

Theoretical continuous particle trajectories are simulated using the so-called leapfrog method. Define the "potential energy" function $U(\vartheta) = -\log(p(\vartheta|\mathbf{Y}))$, where $p(\vartheta|\mathbf{Y})$ is the unstandardized posterior and define $K(p)$ to be the "kinetic energy" function. The kinetic energy function is taken to be proportional to a zero-mean Multivariate Normal, so that $K(p) \propto (p'M^{-1}p)$ (e.g., Neal, 2011), where M is a positive-definite matrix that reflects differential scaling in ϑ, but can also be the identity matrix. The Metropolis acceptance ratio for each proposal is a function of both $U(\vartheta)$ and $K(p)$, which is defined as the Hamiltonian $H(\vartheta,p) = U(\vartheta) + K(p)$. Using $H(\vartheta^*,p^*)$ to denote the value of the Hamiltonian at some proposal ϑ^*, we accept the proposed sample with probability min(1, $exp(H(\vartheta, p) - H(\vartheta^*,p^*))$). HMC proposals are generally very efficient and acceptance probabilities of 0.9 are not uncommon.

HMC explicitly incorporates posterior geometry in its proposals via the gradient function with respect to the parameters $\nabla_\theta H(\vartheta,p) = \nabla_\theta U(\vartheta) = \frac{\partial U}{\partial \vartheta}$, which is computed numerically using the autodiff function. Let (L) be the number of leapfrog steps per trajectory and let (ε) denote the size of each step. Trajectories are simulated by iteratively updating momentum p and position ϑ vectors for $t = 1, \ldots, L$ steps as follows:

$$p_{t+1} = p_t - 0.5\epsilon\nabla_\theta U(\vartheta_t)$$

$$\vartheta_{t+1} = \vartheta_t + \epsilon M p_{t+1}$$

$$p_{t+1} = p_t - 0.5\epsilon\nabla_\theta U(\vartheta_{t+1})$$

Only the value of ϑ at the final position $t = L$ is recorded and used as the proposal in the subsequent Metropolis step. In flat portions of -log(posterior) where $\nabla_\theta U(\vartheta) \approx 0$, neither the momentum nor the position vector are updated, which reflects the frictionless property of Hamiltonian dynamics. This in-part motivates the use of weakly-informative rather than flat or uninformative priors when sampling takes place via NUTS HMC.

The NUTS variant of HMC adapts the length of each trajectory $(L\varepsilon)$ by varying both L and ε during its warm-up phase. If ε is too large, simulated trajectories deviate from the true contours of the posterior and the sampler may fail to explore areas of high curvature. These types of unfavorable trajectories end in so-called "divergent transitions", which bias the resultant MCMC chain. If either ε or L are too small, simulated trajectories will fail to sufficiently move around the posterior and successive samples will be too auto-correlated. Conversely, if L is too large, trajectories will turn back onto themselves, resulting in similarly auto-correlated samples. In the implementation of the adaptive NUTS algorithm featured in Stan, users specify the target acceptance probability using the "adapt delta" parameter and the algorithm stops trajectories from performing U-turns and returning to previously explored areas of the posterior.

B.2. Bayesian model averaging details

Here, we employ Bayesian Model Averaging (BMA) to avoid basing our inference about the treatment effects on a single model. Suppose $M_1, \ldots, M_K$ are the K models considered for a specific response vector Y. For some model parameter of interest (β), the posterior distribution averaged across the K models is (e.g., Hoeting et al., 1999):

$$p(\beta|Y) = \sum_{K} p(\beta|M_k, Y) \times p(M_k|Y)$$

In the equation above, $p(\beta|M_k, Y)$ is the posterior distribution of β in each of the models considered – now formally conditional on the model M_k in addition to Y – and $p(M_k|Y)$ is the "weight" of each model.

There are different ways to compute model weights. A common proposal in statistical literature is:

$$p(M_k|Y) = \frac{p(Y|M_k)p(M_k)}{\sum_K p(Y|M_k)p(M_k)}$$

where $p(Y|M_k) = \int f(Y|\beta, M_k)p(\beta|M_k)d\beta$ is the marginal likelihood, dependent on the prior of β (i.e., $p(\beta|M_k)$) under model M_k. The dependence of the marginal likelihood on the prior, the need to compute the integral for each model considered, and the need to specify prior distributions of the *model* $p(M_k)$ have all been noted as limitations of this approach (e.g., Hoeting et al., 1999). An alternative relies on Akaike-like weights (Akaike, 1978), applied to a Bayesian information criterion, such as the WAIC. These weights are taken as: $w_k = exp(WAIC_k)/\sum_K exp(WAIC_k)$

Yao et al. (2018) recently extend prior work on averaging point estimates to encompass posterior distributions. Model weights are estimated post-model fitting by a solver, for example the optim() solver in R, with optimal weights computed by maximizing the leave-one-out log-score. The authors recommended using their proposed stacking method because the weights are selected with a focus on the predictive performance of the combined posterior, wherein the Akaike-like weights can serve as good initial "guesses" at which to start the solver. The method is implemented in the loo R package (Vehtari et al., 2020), which we used in to perform model averaging in the article.

References

Agresti, A. (2018). *Statistical methods for the social sciences*. Pearson.

Akaike, H. (1978). On the likelihood of a time series model. *Journal of the Royal Statistical Society: Series D (The Statistician), 27*(3–4), 217–235.

Allsopp, D. H., Kyger, M. M., & Lovin, L. H. (2007). *Teaching mathematics meaningfully: Solutions for reaching struggling learners*. ERIC.

Anderson, L. M., Shinn, C., Fullilove, M. T., Scrimshaw, S. C., Fielding, J. E., Normand, J., Carande-Kulis, V. G., & Task Force on Community Preventive Services. (2003). The effectiveness of early childhood development programs: A systematic review. *American Journal of Preventive Medicine, 24*(3), 32–46.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*(4), 561–573.

Baroody, A. (2016). Curricular approaches to connecting subtraction to addition and fostering fluency with basic differences in grade 1. *PNA, 10*(3), 161–190. https://doi.org/10.30827/pna.v10i3.6087.

Baroody, A. J. (2003). The development of adaptive expertise and flexibility: The integration of conceptual and procedural knowledge. In A. J. Baroody, & A. Dowker (Eds.), *The development of arithmetic concepts and skills: Constructing adaptive expertise* (pp. 1–33). Lawrence Erlbaum Associates Publishers.

Baroody, A. J. (2006). Why children have difficulties mastering the basic number combinations and how to help them. *Teaching Children Mathematics, 13*(1), 22–31.

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software, 80*, 1–28.

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science, 2*(1), 77–101.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research, 33*(2), 261–304.

Cai, J., & Brook, M. (2006). Looking back in problem solving. *Mathematics Teaching Incorporating Micromath, 196*, 42–45.

Camilli, G., Vargas, S., Ryan, S., & Barnett, W. S. (2010). Meta-analysis of the effects of early education interventions on cognitive and social development. *Teachers College Record, 112*(3), 579–620.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., & Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1), 1–32.

Carpenter, T. P. (1985). Learning to add and subtract: An exercise in problem solving. *Teaching and Learning Mathematical Problem Solving: Multiple Research Perspectives, 17*, 40.

Carpenter, T. P., Fennema, E., Franke, M. L., Levi, L., & Empson, S. B. (2015). *Children's mathematics: Cognitively guided instruction* (2nd ed.). Heinemann.

Carpenter, T. P., Franke, M. L., Jacobs, V. R., Fennema, E., & Empson, S. B. (1998). A longitudinal study of invention and understanding in children's multidigit addition and subtraction. *Journal for Research in Mathematics Education, 29*(1), 3–20. https://doi.org/10.2307/749715.

Carr, M., & Alexeev, N. (2011). Fluency, accuracy, and gender predict developmental trajectories of arithmetic strategies. *Journal of Educational Psychology, 103*(3), 617–631. https://doi.org/10.1037/a0023864.

Carr, M., & Davis, H. (2001). Gender differences in arithmetic strategy use: A function of skill and preference. *Contemporary Educational Psychology, 26*(3), 330–347. https://doi.org/10.1006/ceps.2000.1059.

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. *Economics of Education Review, 28*(4), 415–427. https://doi.org/10.1016/j.econedurev.2008.09.003.

Claessens, A., & Engel, M. (2013). How important is where you start? Early mathematics knowledge and later school success. *Teachers College Record, 115*(6), 1–29.

Clements, D. H., Dumas, D., Dong, Y., Banse, H. W., Sarama, J., & Day-Hess, C. A. (2020). Strategy diversity in early mathematics classrooms. *Contemporary Educational Psychology, 60*, Article 101834.

Clements, D. H., & Sarama, J. (2021). *Learning and teaching early math: The learning trajectories approach*. Routledge & CRC Press.

Clements, D. H., Sarama, J., Baroody, A. J., Kutaka, T. S., Chernyavskiy, P., Joswick, C., Cong, M., & Joseph, E. (2021). Comparing the efficacy of early arithmetic instruction based on a learning trajectory and teaching-to-a-target. *Journal of Educational Psychology, 113*(7), 1323.

Clements, D. H., Sarama, J., Wolfe, C. B., & Spitler, M. E. (2013). Longitudinal evaluation of a scale-up model for teaching mathematics with trajectories and technologies: Persistence of effects in the third year. *American Educational Research Journal, 50*(4), 812–850. https://doi.org/10.3102/0002831212469270.

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based early Maths assessment. *Educational Psychology, 28*(4), 457–482. https://doi.org/10.1080/01443410701777272.

Council, N.R. (2000). *Inquiry and the national science education standards: A guide for teaching and learning*. National Academies Press.

Daro, P., Mosher, F. A., & Corcoran, T. B. (2011). *Learning trajectories in mathematics: A foundation for standards, curriculum, assessment, and instruction*. CPRE Research Report# RR-68. Consortium for Policy Research in Education

Draper, D., Hodges, J. S., Leamer, E. E., Morris, C. N., & Rubin, D. B. (1987). *A research agenda for assessment and propagation of model uncertainty*. Report N-2683-RC. Rand Corporation

Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement. *Developmental Psychology, 43*(6), 1428–1446. https://doi.org/10.1037/0012-1649.43.6.1428.

Fennema, E., Carpenter, T. P., Jacobs, V. R., Franke, M. L., & Levi, L. W. (1998). A longitudinal study of gender differences in young children's mathematical thinking. *Educational Researcher, 27*(5), 6–11. https://doi.org/10.2307/1176733.

Flavell, J. H. (1970). Developmental studies of mediated memory. In H. W. Reese, & L. P. Lipsitt (Eds.), *Vol. 5. Advances in child development and behavior* (pp. 181–211). JAI. https://doi.org/10.1016/S0065-2407(08)60467-X.

Frye, D., Baroody, A. J., Burchinal, M., Carver, S. M., Jordan, N. C., & McDowell, J. (2013). Teaching math to young children. In *What Works Clearinghouse*. https://eric.ed.gov/?id=ED544376.

Fuchs, L. S., Compton, D. L., Fuchs, D., Paulsen, K., Bryant, J. D., & Hamlett, C. L. (2005). The prevention, identification, and cognitive determinants of math difficulty. *Journal of Educational Psychology, 97*(3), 493–513.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2012). The early prevention of mathematics difficulty: Its power and limitations. *Journal of Learning Disabilities, 45*(3), 257–269.

Fuchs, L. S., Powell, S. R., Fall, A.-M., Roberts, G., Cirino, P., Fuchs, D., & Gilbert, J. K. (2020). Do the processes engaged during mathematical word-problem solving differ along the distribution of word-problem competence? *Contemporary Educational Psychology, 60*, 101811.

Fuchs, L. S., & Vaughn, S. (2012). Responsiveness-to-intervention: A decade later. *Journal of Learning Disabilities, 45*(3), 195–203. https://doi.org/10.1177/0022219412442150.

Fullerton, A. S., & Xu, J. (2016). *Ordered regression models: Parallel, partial, and non-parallel alternatives*. CRC Press.

Fuson, K. C. (1992). Relationships between counting and cardinality from age 2 to age 8. In J. Bideaud, C. Meljac, & J.-P. Fischer (Eds.), *Pathways to number: Children's developing numerical abilities* (pp. 127–149). Lawrence Erlbaum Associates, Inc.

Fuson, K. C. (2012). *Children's counting and concepts of number*. Springer Science & Business Media.

Garon-Carrier, G., Boivin, M., Lemelin, J.-P., Kovas, Y., Parent, S., Séguin, J. R., Vitaro, F., Tremblay, R. E., & Dionne, G. (2018). Early developmental trajectories of number knowledge and math achievement from 4 to 10 years: Low-persistent profile and early-life predictors. *Journal of School Psychology, 68*, 84–98.

Geary, D. C. (2011). Cognitive predictors of achievement growth in mathematics: A 5-year longitudinal study. *Developmental Psychology, 47*(6), 1539–1552.

Geeslin, W., & Graham, K. (1992). *Proceedings of the conference of the International Group for the Psychology of mathematics education (16th, Durham, NH, august 6–11, 1992)* (Vol. I-III).

Gelman, A., & Hennig, C. (2017). Beyond subjective and objective in statistics. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 180*(4), 967–1033.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*(4), 457–472. https://doi.org/10.1214/ss/1177011136.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy, 19*(10), 555.

Gelman, R., & Gallistel, C. R. (1986). *The child's understanding of number*. Harvard University Press.

Gelman, R., & Meck, E. (1983). Preschoolers' counting: Principles before skill. *Cognition, 13*(3), 343–359.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability* (3rd ed.). Pro-Ed.

Ginsburg, H. P., Lee, J. S., & Boyd, J. S. (2008). Mathematics education for young children: What it is and how to promote it. *Social Policy Report, 22*(1), 1–24. https://doi.org/10.1002/j.2379-3988.2008.tb00054.x.

Greer, B. (1997). Modelling reality in mathematics classrooms: The case of word problems. *Learning and Instruction, 7*(4), 293–307.

Hodges, J. S. (1987). Uncertainty, policy analysis and statistics. *Statistical Science, 2*(3), 259–275.

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with comments by M. Clyde, David Draper and EI George, and a rejoinder by the authors). *Statistical Science, 14*(4), 382–417.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15*(1), 1593–1623.

Hougaard, P. (1999). Multi-state models: A review. *Lifetime Data Analysis, 5*(3), 239–264.

Jackson, C. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software, 38*, 1–28. https://doi.org/10.18637/jss.v038.i08.

Jordan, N. C., Glutting, J., & Ramineni, C. (2010). The importance of number sense to mathematics achievement in first and third grades. *Learning and Individual Differences, 20*(2), 82–88. https://doi.org/10.1016/j.lindif.2009.07.004.

Jordan, N. C., Kaplan, D., Ramineni, C., & Locuniak, M. N. (2009). Early math matters: Kindergarten number competence and later mathematics outcomes. *Developmental Psychology, 45*(3), 850–867. https://doi.org/10.1037/a0014939.

Kerkman, D. D., & Siegler, R. S. (1993). Individual differences and adaptive flexibility in lower-income children's strategy choices. *Learning and Individual Differences, 5*(2), 113–136.

Kieren, T., & Pirie, S. (1992). The answer determines the question: Interventions and the growth of mathematical understanding. In *In Geeslin and Graham (Eds.)*. Durham, NH: Proceedings of the Sixteenth Psychology of Mathematics Education Conference.

Kutaka, T. S., Ren, L., Smith, W. M., Beattie, H. L., Edwards, C. P., Green, J. L., … Lewis, W. J. (2018). Examining change in K-3 teachers' mathematical knowledge, attitudes, and beliefs: The case of primarily math. *Journal of Mathematics Teacher Education, 21*(2), 147–177. https://doi.org/10.1007/s10857-016-9355-x.

Kutaka, T. S., Smith, W. M., Albano, A. D., Edwards, C. P., Ren, L., Beattie, H. L., Lewis, W. J., Heaton, R. M., & Stroup, W. W. (2017). Connecting teacher professional development and student mathematics achievement: A 4-year study of an elementary mathematics specialist program. *Journal of Teacher Education, 68*(2), 140–154. https://doi.org/10.1177/0022487116687551.

Lembke, E. S., Hampton, D., & Beyers, S. J. (2012). Response to intervention in mathematics: Critical elements. *Psychology in the Schools, 49*(3), 257–272.

Lesh, R., Hoover, M., Hole, B., Kelly, A., & Post, T. R. (2000). Principles for developing thought-revealing activities for students and teachers. In A. Kelly, & R. Lesh (Eds.), *Research design in mathematics and science education* (pp. 591–646). Lawrence Erlbaum Associates.

Lesh, R. A., & Lamon, S. J. (2013). *Assessment of authentic performance in school mathematics*. Routledge.

Lobato, J., & Walter, C. D. (2017). *A taxonomy of approaches to learning trajectories and progressions*. The National Council of Teachers of Mathematics: The compendium for research in mathematics education.

Maloney, A. P., Confrey, J., & Nguyen, K. H. (2014). *Learning over time: Learning trajectories in mathematics education*. Information Age Publishing, INC.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B (Methodology), 42*(2), 109–142.

McElreath, R., & Koster, J. (2014). Using multilevel models to estimate variation in foraging returns. *Human Nature, 25*(1), 100–120.

McLeod, B. M. (1992). Research on affect in mathematics education: A reconceptualization. In D. Grouws (Ed.), *Handbook of research on mathematics teaching and learning* (pp. 575–596). Information Age Publishing.

Monnahan, C. C., Thorson, J. T., & Branch, T. A. (2017). Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution, 8*(3), 339–348.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. https://www.nctm.org/uploadedFiles/Standards_and_Positions/PSSM_ExecutiveSummary.pdf.

National Council of Teachers of Mathematics. (2014). Principles to actions. https://www.nctm.org/uploadedFiles/Standards_and_Positions/PtAExecutiveSummary.pdf.

National Governors Association & Council of Chief State School Officers. (2010). Mathematics Standards | Common Core State Standards Initiative. https://learning.ccsso.org/wp-content/uploads/2022/11/ADA-Compliant-Math-Standards.pdf.

National Mathematics Advisory Panel. (2008). Foundations for success: The final report of the National Mathematics Advisory Panel. https://files.eric.ed.gov/fulltext/ED500486.pdf.

National Research Council. (2001). In J. Kilpatrick, J. Swafford, & B. Findell (Eds.), *Adding it up: Helping children learn mathematics*. The National Academies Press. https://doi.org/10.17226/9822.
National Research Council. (2009). *Mathematics learning in early childhood: Paths toward excellence and equity*. The National Academies Press. https://doi.org/10.17226/12519.
Naus, M. J., & Ornstein, P. A. (1983). Development of memory strategies: Analysis, questions, and issues. *Trends in Memory Development Research, 9*, 1–30. https://doi.org/10.1159/000407963.
Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In S. Brooks, A. Gelman, G. Jones, & X. Meng (Eds.), *Handbook of Markov chain Monte Carlo* (pp. 112–172). Chapman & Hall Press.
Nishio, M., & Arakawa, A. (2019). Performance of Hamiltonian Monte Carlo and no-U-turn sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution, 51*(1), 1–12.
Pagani, L., Fitzpatrick, C., Archambault, I., & Janosz, M. (2010). School readiness and later achievement: A French Canadian replication and extension. *Developmental Psychology, 46*, 984–994. https://doi.org/10.1037/a0018881.
Peyhardi, J., Trottier, C., & Guédon, Y. (2015). A new specification of generalized linear models for categorical responses. *Biometrika, 102*(4), 889–906.
Polya, G. (1985). *How to solve it (2nd ed)*. Princeton University Press.
Raman, R., & Hedeker, D. (2005). A mixed-effects regression model for three-level ordinal response data. *Statistics in Medicine, 24*(21), 3331–3345. https://doi.org/10.1002/sim.2186.
Rittle-Johnson, B., Siegler, R. S., & Alibali, M. W. (2001). Developing conceptual understanding and procedural skill in mathematics: An iterative process. *Journal of Educational Psychology, 93*(2), 346–362. https://doi.org/10.1037/0022-0663.93.2.346.
Rodrigues, J. (2021). Get more eyes on your work: Visual approaches for dissemination and translation of education research. *Educational Researcher, 50*(9), 657–663. https://doi.org/10.3102/0013189X211035351.
Samejima, F., van der Liden, W. J., & Hambleton, R. (1997). *Handbook of modern item response theory*. Springer.
Sarama, J., & Clements, D. H. (2009). *Early childhood mathematics education research: Learning trajectories for young children*. Routledge. https://doi.org/10.4324/9780203883785.
Sarama, J., & Clements, D. H. (2019). The building blocks and TRIAD projects. In P. Sztajn, & H. Wilson (Eds.), *Learning trajectories for teachers: Designing effective professional development for math instruction* (pp. 104–131). Teachers College Press.
Sarama, J., Clements, D. H., Baroody, A. J., Kutaka, T. S., Chernyavskiy, P., Shi, J., & Cong, M. (2021). Testing a theoretical assumption of a learning-trajectories approach in teaching length measurement to kindergartners. *AERA Open, 7*. https://doi.org/10.1177/23328584211026657.
Sarnecka, B. W., & Carey, S. (2008). How counting represents number: What children must learn and when they learn it. *Cognition, 108*(3), 662–674. https://doi.org/10.1016/j.cognition.2008.05.007.
Siegler, R. S. (1987). Some general conclusions about children's strategy choice procedures. *International Journal of Psychology, 22*, 729–749.
Siegler, R. S. (1991). Strategy choice and strategy discovery. *Learning and Instruction, 1*(1), 89–102. https://doi.org/10.1016/0959-4752(91)90020-9.
Siegler, R. S. (1995). How does change occur: A microgenetic study of number conservation. *Cognitive Psychology, 28*(3), 225–273.
Siegler, R. S. (2007). Cognitive variability. *Developmental Science, 10*(1), 104–109.
Siegler, R. S., & Jenkins, E. (1989). *How children discover new strategies* (1st ed.). Psychology Press.
Simon, M. A. (1995). Reconstructing mathematics pedagogy from a constructivist perspective. *Journal for Research in Mathematics Education, 26*(2), 114–145. https://doi.org/10.2307/749205.
Smithson, M., & Merkle, E. C. (2013). *Generalized linear models for categorical and continuous limited dependent variables*. CRC Press.
Snijders, T. A. B., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. SAGE.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 485–493.
Supovitz, J. A., Ebby, C. B., Remillard, J., & Nathenson, R. A. (2018). Experimental impacts of the ongoing assessment project on teachers and students. Consortium for policy research in education (CPRE). *Research Reports*. Retrieved from https://repository.upenn.edu/cpre_researchreports/107.
Supovitz, J. A., Ebby, C. B., Remillard, J. T., & Nathenson, R. (2021). Experimental impacts of learning trajectory–oriented formative assessment on student problem-solving accuracy and strategy sophistication. *Journal for Research in Mathematics Education, 52*(4), 444–475. https://doi.org/10.5951/jresematheduc-2021-0032.
Tanner-Smith, E. E., Durlak, J. A., & Marx, R. A. (2018). Empirically based mean effect size distributions for universal prevention programs targeting school-aged youth: A review of meta-analyses. *Prevention Science, 19*(8), 1091–1101.
Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis, 11*(3), 275–295.
Tutz, G. (2005). Modelling of repeated ordered measurements by isotonic sequential regression. *Statistical Modelling, 5*(4), 269–287.
Tutz, G., & Groll, A. (2013). Likelihood-based boosting in binary and ordinal random effects models. *Journal of Computational and Graphical Statistics, 22*(2), 356–378.
VanDerHeyden, A., & Harvey, M. (2013). Using data to advance learning outcomes in schools. *Journal of Positive Behavior Interventions, 15*(4), 205–213.
VanDerHeyden, A. M., Broussard, C., Snyder, P., George, J., Lafleur, S. M., & Williams, C. (2011). Measurement of kindergartners' understanding of early mathematical concepts. *School Psychology Review, 40*(2), 296–306.
Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., Gelman, A., Goodrich, B., Piironen, J., & Nicenboim, B. (2020). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models (2.4.1). https://CRAN.R-project.org/package=loo.
Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing, 27*(5), 1413–1432.
Verschaffel, L., Greer, B., & De Corte, E. (2000). Making sense of word problems. *Lisse, The Netherlands, 224*, 224.
Watanabe, S. (2013). WAIC and WBIC are information criteria for singular statistical model evaluation. *Proceedings of the Workshop on Information Theoretic Methods in Science and Engineering*, 90–94.
Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P. E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*(7), 352–360. https://doi.org/10.3102/0013189X14553660.
Yao, Y., Vehtari, A., Simpson, D., & Gelman, A. (2018). Using stacking to average Bayesian predictive distributions (with discussion). *Bayesian Analysis, 13*(3), 917–1007.
Zeger, S. L., Liang, K., & Albert, P. (1988). *Liang KY ja Albert PS (1988): Models for Longitudinal Data: a GEE Approach. 44* pp. 1049–1060).
Zhu, Z. (2007). Gender differences in mathematical problem solving patterns: A review of literature. *International Education Journal, 8*(2), 17.