

Children's Cognitive Reflection Predicts Conceptual Understanding in Science and Mathematics



Andrew G. Young¹  and Andrew Shtulman²

¹Department of Psychology, Northeastern Illinois University, and ²Department of Psychology, Occidental College

Psychological Science
2020, Vol. 31(11) 1396–1408
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0956797620954449
www.psychologicalscience.org/PS



Abstract

The Cognitive Reflection Test (CRT) is a widely used measure of adults' propensity to engage in reflective analytic thought. The CRT is strongly predictive of many diverse psychological factors but unsuitable for use with developmental samples. Here, we examined a children's CRT, the CRT–Developmental (CRT–D), and investigated its predictive utility in the domains of science and mathematics. School-age children ($N = 152$) completed the CRT–D, measures of executive functioning, measures of rational thinking, and measures of vitalist-biology and mathematical-equivalence concepts. CRT–D performance predicted conceptual understanding in both domains after we adjusted for children's age, executive functioning, and rational thinking. These findings suggest that cognitive reflection supports conceptual knowledge in early science and mathematics and, moreover, demonstrate the theoretical and practical importance of children's cognitive reflection. The CRT–D will allow researchers to investigate the development, malleability, and consequences of children's cognitive reflection.

Keywords

cognitive reflection, cognitive development, conceptual change, folk biology, mathematical equivalence, open data, open materials

Received 8/24/18; Revision accepted 6/1/20

The Cognitive Reflection Test (CRT; Frederick, 2005) is the dominant measure of adult individual differences in analytic versus intuitive thinking. The test was designed to measure a person's tendency to override an intuitive response that is incorrect and engage in reflection that leads to a correct response. Consider the well-known bat-and-ball item: "A bat and a ball cost \$1.10 in total. The bat costs \$1 more than the ball. How much does the ball cost?" Many adults provide the intuitively triggered response of 10 cents, defaulting to subtraction. However, the correct answer is 5 cents (the bat must cost \$1.05 for their sum to be \$1.10 and their difference to be \$1.00), and adults who provide that answer have presumably engaged in analytic reflection, realizing that the intuitive response was incorrect and generating a correct response in its place.

Performance on the CRT is a powerful predictor of rational thinking, including normative decision-making on heuristics-and-biases tasks and normative thinking

dispositions (Frederick, 2005; Stanovich, West, & Toplak, 2016; Toplak, West, & Stanovich, 2011). CRT scores also predict science understanding (Shtulman & McCallum, 2014), science acceptance (Gervais, 2015), rejection of religious and paranormal ideas (Pennycook, Fugelsang, & Koehler, 2015), utilitarian moral reasoning (Royzman, Landy, & Leeman, 2015), causal learning (Don, Goldwater, Otto, & Livesey, 2016), fake-news detection (Pennycook & Rand, 2019), cooperation (Corgnet, Espín, & Hernández-González, 2015), and avoidance of stereotyping (Hammond & Cimpian, 2017). The CRT has thus garnered broad utility and interest.

The present research extends the study of cognitive reflection to children. Floor effects in adolescent and

Corresponding Author:

Andrew G. Young, Northeastern Illinois University, Department of Psychology
E-mail: ayoung20@neiu.edu

certain adult populations as well as heavy reliance on numeracy suggest that the original CRT is poorly suited for child samples (Primi, Morsanyi, Chiesi, Donati, & Hamilton, 2016; Thomson & Oppenheimer, 2016). As a result, children's cognitive reflection has not been studied. A CRT for children could be used to investigate the developmental trajectory of cognitive reflection as well as its malleability with experience or intervention. Similarly, a developmental measure might help adjudicate between competing accounts of cognitive reflection in adults (e.g., Stupple, Pitchford, Ball, Hunt, & Steel, 2017; Szaszi, Szollosi, Palfi, & Aczel, 2017; Travers, Rolison, & Feeney, 2016). We directly considered another contribution: examining the development of important domains of thought, such as scientific and mathematical reasoning.

We recently developed a CRT for elementary-school-age children, the CRT-Developmental (CRT-D), which consists of nine items similar in structure to the original CRT items. Each item has a highly available intuitive (but incorrect) lure and a correct response that we expected school-age children to be capable of producing after reflection or analytic thinking. A preliminary investigation assessed whether children's and adults' CRT-D scores predicted performance on a battery of heuristics-and-biases tasks and measures of normative thinking (Young, Powers, Pilgrim, & Shtulman, 2018). For adults, the CRT-D correlated strongly with the original CRT and yielded similar correlations with heuristics-and-biases tasks. For children, the CRT-D predicted performance on heuristics-and-biases tasks as well as normative thinking dispositions, even after analyses adjusted for age. These results provide preliminary support for the CRT-D as a valid measure of children's cognitive reflection. However, the productivity of this construct will depend on its ability to predict reasoning beyond the domain of rational thinking and independently of other domain-general cognitive abilities.

The present study focused on the predictive utility of the CRT-D in science and mathematics, domains characterized by conceptual change. Research with adults suggests that cognitive reflection supports conceptual change. Shtulman and McCallum (2014) found that the CRT explained more variance in college students' achievement of conceptual change in six domains of science (astronomy, evolution, geology, mechanics, perception, and thermodynamics) than did science and mathematics coursework; statistical reasoning ability; and nature-of-science understanding combined. Similarly, CRT performance predicts students' understanding of various secondary school math concepts (Gómez-Chacón, García-Madruga, Vila, Elosúa, & Rodríguez, 2014). Here, we asked the developmentally parallel question of

whether the CRT-D predicts school-age children's conceptual understanding of vitalist biology and mathematical equivalence.

Vitalist biology is a culturally widespread theory of life that develops from ages 5 to 12 years (Inagaki & Hatano, 2002). According to this theory, living organisms are systems that use vital substances (e.g., food, air, and water) to produce energy and sustain life, health, and growth. Prior to gaining awareness of vitalist biology, children conceive of life in terms of agency and animism. For example, young children may judge nonliving animate entities (e.g., the sun, an airplane) as alive and living inanimate entities (e.g., plants) as not alive (Zaitchik, Iqbal, & Carey, 2014). In addition, children lacking a mature theory of vitalist biology often report that body parts have single independent functions (e.g., the stomach is for storing food), failing to conceive of those functions as interrelated and life sustaining (Slaughter & Lyons, 2003).

Mathematical equivalence is the principle that two sides of an equation are interchangeable and represent the same value. Understanding of mathematical equivalence promotes math achievement (McNeil, Hornburg, Devlin, Carrazza, & McKeever, 2019) and is foundational to formal algebra (Knuth, Stephens, McNeil, & Alibali, 2006). However, narrow experience with standard-format equations (e.g., $2 + 7 = \underline{\quad}$) leads many children to an entrenched misconception of the equal sign as an operational symbol (i.e., put the answer or add all the numbers) rather than a relational symbol (McNeil & Alibali, 2005). As a consequence, 60% to 88% of children in the United States between the ages of 7 and 11 years solve problems with operations on both sides of the equal sign (e.g., $2 + 7 = 6 + \underline{\quad}$) incorrectly, usually by adding all the numbers or adding all the numbers before the equal sign (Hornburg, Wang, & McNeil, 2018).

In this study, we measured school-age children's performance on the CRT-D and understanding of vitalist biology and mathematical equivalence. We additionally measured children's executive functions, including exogenous and endogenous set shifting, inhibitory control, and working memory. Research with adults suggests that the CRT's predictive strength is largely independent of executive functions (Toplak et al., 2011), although this may not be true of children. Further, executive functions play a critical role in children's construction of vitalist biology (Bascandziev, Tardiff, Zaitchik, & Carey, 2018; Zaitchik et al., 2014) and development of mathematics proficiency (Cragg & Gilmore, 2014). Finally, we measured children's performance on a small number of rational-thinking tasks, measuring their reliance on heuristics and biases and their disposition toward normative-thinking strategies because such

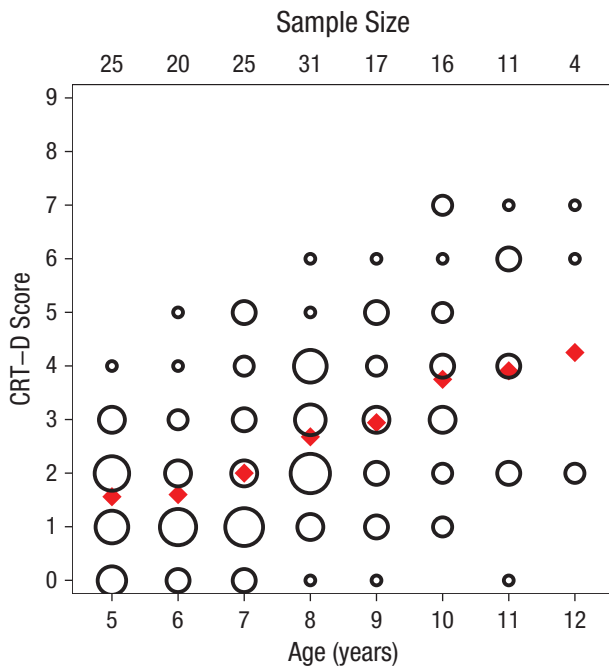


Fig. 1. Distribution of Cognitive Reflection Test-Developmental (CRT-D) scores by age. Red diamonds represent means, and the size of each circle is proportional to the number of children with the given score. The sample size at each age is shown above the graph. One 58-month-old child and two other children with missing age data are not represented (CRT-D scores = 0, 2, and 4, respectively).

tasks are also implicated in science and mathematics reasoning (e.g., Sinatra, Southerland, McConaughy, & Demastes, 2003; Wong, 2018). We asked whether the CRT-D is a useful predictor of vitalist-biology and mathematical-equivalence understanding after adjusting for children's age, executive functions, and rational thinking.

Method

Participants

One hundred fifty-two elementary school children participated ($M = 8$ years, 2 months; $SD = 1$ year, 11 months; 85 female, 67 male). We recruited children between the ages of 5 and 12 years who were in kindergarten through sixth grade (see Fig. 1 for the age distribution). We recruited children at public playgrounds and a children's museum during two academic terms, using the end of the second term as the predetermined stopping point for data collection. Thus, the sample size was determined by the practicalities of participant recruitment. The final sample is 58% larger than the preliminary study of the CRT-D (Young et al., 2018), which generally found moderate to large correlations between the CRT-D and other measures (median $r = .30$). This

sample size was sufficient for examining the predictive utility of the CRT-D given our analytic approach (i.e., Bayesian estimation, regularizing priors, and cross-validation; see below).

Measures

Study materials are available on OSF at <https://osf.io/e72ka/>.

CRT-Developmental. Children answered nine child-friendly cognitive-reflection items similar in structure to those in the original CRT for adults (Table 1). Three items (Questions 1–3) were adapted from the nonnumerical CRT developed by Thomson and Oppenheimer (2016). The remaining items were found by searching for children's "brain teasers" online or developed by the researchers. We used the number of correct responses as children's score; higher scores indicated greater cognitive reflection. The McDonald's ω_{total} of the scale was .77, and Cronbach's α was .56. McDonald's ω_{total} is an estimate of the total reliability of a test. It is a less biased estimate of reliability than Cronbach's α in most circumstances and equivalent to Cronbach's α when the latter's often unrealistic assumptions are met (Zinbarg, Revelle, Yovel, & Li, 2005).

Executive-function tasks.

Toolbox Dimensional Change Card Sort. Children completed the tablet-based Dimensional Change Card Sort (DCCS) from the NIH Toolbox Cognition Battery (Zelazo et al., 2013; iPad Version 1.11). This test measures cognitive flexibility in the context of exogenously cued set shifting, requiring children to match multidimensional pictures first by one dimension (e.g., shape) and then by another (e.g., color). The Toolbox DCCS consists of four blocks: practice (four trials for each dimension), preswitch (five trials for the first dimension), postswitch (five trials for the second dimension), and mixed (30 trials shifting between dimensions). Scoring of the Toolbox DCCS is based on both accuracy and reaction time (Zelazo et al., 2013). We used uncorrected standardized scores ($M = 100$, $SD = 15$), which reflect the overall level of performance relative to the entire NIH Toolbox normative sample, regardless of age or other demographic factors (test-retest intraclass correlation coefficient [ICC] = .92). Higher scores indicate better performance.

Toolbox Flanker Inhibitory Control and Attention Test. Children completed the tablet-based Flanker Test from the NIH Toolbox Cognition Battery (Zelazo et al., 2013; iPad Version 1.11). The test measures both attention and inhibitory control, requiring children to indicate the left-right orientation of a middle stimulus while inhibiting

Table 1. Cognitive Reflection Test–Developmental (CRT–D) Questions and Percentages of Response Types

Question	Correct	Intuitive incorrect	Other incorrect
1. If you’re running a race and you pass the person in second place, what place are you in? (correct: second; intuitive: first)	17.8	70.4	11.8
2. Emily’s father has three daughters. The first two are named Monday and Tuesday. What is the third daughter’s name? (correct: Emily; intuitive: Wednesday or other days of the week)	7.9	87.5	4.6
3. A farmer has 5 sheep, all but 3 run away. How many are left? (correct: three; intuitive: two)	20.4	65.8	13.8
4. If there are 3 apples and you take away 2, how many do you have? (correct: two; intuitive: one)	50.0	40.8	9.2
5. What do cows drink? (correct: water; intuitive: milk)	37.5	61.2	1.3
6. What weighs more, a pound of rocks or a pound of feathers? (correct: same weight; intuitive: rocks)	7.9	85.5	6.6
7. What hatches from a butterfly egg? (correct: caterpillar/larva; intuitive: baby butterfly)	63.8	24.3	11.8
8. Who makes Christmas presents at the North Pole? (correct: elves, parents, or no one; intuitive: Santa)	25.0	73.7	1.3
9. Anna is playing foursquare with her three friends: Eeny, Meeny, and Miny. Who is the fourth player? (correct: Anna; intuitive: Mo)	19.7	66.4	13.8

attention to four flanking stimuli. On congruent trials, all stimuli are pointing in the same direction, whereas on incongruent trials, the flanking stimuli are pointing in the opposite direction from the middle stimulus. Children 8 years of age and older completed four practice trials and 20 test trials with arrows as the stimuli. Children younger than 8 years completed the same number of trials with fish as stimuli. Younger children who missed no more than one congruent and one incongruent trial also completed the arrow test block. Scoring of the Toolbox Flanker is based on both accuracy and reaction time (Zelazo et al., 2013). As with the Toolbox DCCS, we used uncorrected standardized scores based on the NIH Toolbox normative sample ($M = 100$, $SD = 15$; test-retest $ICC = .92$).

Verbal fluency. Children completed a verbal fluency task, in which they named as many animals as they could without repetition in 1 min (Bascandziev et al., 2018; Munakata, Snyder, & Chatham, 2012). This task requires identifying relevant semantic categories, accessing the items within those categories, and switching to new categories when needed. To be successful, children must detect the need to switch (e.g., when they cannot think of more pets) and select what to switch to (e.g., farm animals or zoo animals). This task is considered a measure of self-directed cognitive flexibility, requiring endogenous set shifting and proactive executive planning (Munakata et al., 2012). Children’s responses were audio-recorded and transcribed. We used the total number of unique animals generated as children’s score (Bascandziev et al.,

2018). The first author and a research assistant scored each transcription (there was no interrater disagreement).

Backward digit span. Children completed a backward-digit-span task that required both maintenance and manipulation of items in working memory (adapted from the study by Alloway, Gathercole, Kirkwood, & Elliott, 2009). The experimenter read a sequence of numbers at a pace of one per second. Children were then asked to repeat the numbers in reverse order. Children were given a practice trial of three digits and then test trials starting at two digits, increasing by one digit after every two trials. The task ended when children failed both trials of a given length or at the conclusion of the eight-digit trials. We used the highest span with at least one correct trial as children’s score (test-retest reliability = .86; Alloway et al., 2009). Scores could range from 1 to 8 (a score of 1 was assigned if a child failed both two-digit trials).

Heuristics-and-biases tasks.

Denominator neglect. Children completed a tablet-based denominator-neglect task adapted from the study by Toplak, West, and Stanovich (2014). This task measures probabilistic reasoning and miserly information processing in the context of whether children attend to the absolute number of a particular kind of outcome (the numerator) without considering the total number of possible events (the denominator). Children were shown trays of black marbles and white marbles and told that black marbles were the winners (i.e., that the goal of the

game was to select a black marble at random). They were then asked to choose between (a) a smaller tray that contained fewer winning marbles but a higher probability of winning and (b) a larger tray with more winning marbles and a lower probability of winning (e.g., 1:9 vs. 9:91). Children completed six trials of varying ratios, along with three filler trials in which the larger tray also had a higher probability of winning (Stanovich et al., 2016). We used the number of highest probability selections as children's score; higher scores indicated more resistance to denominator neglect (Stanovich et al., 2016). The ω_{total} of the task was .84, and Cronbach's α was .84.

Base-rate sensitivity. Children completed a base-rate-sensitivity task composed of five causal base-rate problems for children (Toplak et al., 2014). These problems measure the tendency to rely on large-sample or expert-provided statistical evidence over concrete personal information (Stanovich et al., 2016). An example scenario is as follows:

Erica wants to go to a baseball game to try to catch a fly ball. She calls the main office and learns that almost all fly balls have been caught in section 43. Just before she chooses her seats, she learns that her friend Jimmy caught a fly ball last week sitting in section 10. Which section is most likely to give Erica the best chance to catch a fly ball?

There were four response choices for each problem: two responses supporting the correct statistical/aggregate choice (e.g., definitely section 43 or probably section 43) and two responses supporting the incorrect concrete/personal choice (e.g., definitely section 10 or probably section 10). We used the total number of statistical/aggregate responses as children's score. The ω_{total} of the task was .62, and Cronbach's α was .44.

Thinking-disposition scales.

Need for Cognition. Children completed a Need for Cognition scale developed and validated for children (Keller et al., 2019). This scale measures children's tendency to engage in and enjoy effortful cognitive activities. Examples from the 14-item child scale are "Thinking is fun for me" and "I like learning new things." Children responded on a 4-point agreement scale (1 = *really disagree* to 4 = *really agree*). We used children's average rating as their score; higher values indicated greater motive to engage in effortful cognitive activities (Keller et al., 2019). The ω_{total} of the scale was .87, and Cronbach's α was .85.

Actively open-minded thinking. Children responded to a modified version of a seven-item scale assessing

actively open-minded thinking (Haran, Ritov, & Mellers, 2013). This scale measures the tendency to weigh new evidence against a favored belief and to consider the opinions of other people in forming one's own. We modified items to be child friendly. However, internal reliability for the scale was very poor ($\omega_{\text{total}} = .45$ and Cronbach's $\alpha = .03$), so we did not consider it for further analyses.

Vitalist-biology understanding.

Body-parts knowledge. Children completed a shortened version of the Body Parts Interview (Bascandzic et al., 2018; Zaitchik et al., 2014). We asked about the function of five body parts: the brain, heart, lungs, stomach, and blood (e.g., "What is the brain for?"). Children's responses were audio-recorded and transcribed. We scored responses using a coding scheme from prior research (Bascandzic et al., 2018; Zaitchik et al., 2014). For a given body part, children could receive 0 to 3 points: 1 point for knowing the organ's function, 1 point for relating the organ's function to another bodily function or biological goal, and 1 point for mentioning that the organ is needed to stay alive. Scores could thus range from 0 to 15; higher scores indicated greater body-parts knowledge. Both authors scored the transcriptions. Interrater reliability was high (94.8% agreement for individual body-part codes; ICC = .91 for total scores).

Living-things knowledge. Children completed the living-things judgment task (Bascandzic et al., 2018; Zaitchik et al., 2014). Children were asked to make judgments (e.g., "Is an X alive; is it a living thing?") for 20 entities from four categories (animals, plants, natural phenomena, and artifacts). We scored the task in terms of completely correct categories. That is, if a child misjudged the living status of any entity within a category, the category was scored as incorrect. Our analyses considered both the number of correct categories and the accuracy of categories individually.

Vitalist-biology composite. Following prior research (Bascandzic et al., 2018; Zaitchik et al., 2014), we created a composite vitalist-biology score by averaging z scores for the body-parts-knowledge measure and living-things-knowledge measure (number of correct categories).

Mathematical equivalence. Children in the second grade and above solved four mathematical-equivalence problems (adapted from the study by McNeil et al., 2019) with operations on both sides of the equal sign ($1 + 5 = _ + 2$; $2 + 7 = 6 + _$; $7 + 1 + 4 = _ + 4$; $3 + 5 + 6 = 3 + _$). We scored problems as correct or incorrect. The ω_{total} of the test was .91, and Cronbach's α was .90. Following prior research, we coded children who solved any of the

problems correctly as demonstrating some understanding of mathematical equivalence because a major shift in conceptual understanding is required to solve just one problem correctly (McNeil et al., 2019).

Procedure

Children completed the study one on one with trained research assistants on-site. Depending on the task and children's ability to read independently, research assistants either read each question aloud or had children read it themselves. Children responded verbally or via iPad touch screen depending on the task.

Children completed the tasks described above in the following order: verbal fluency, body-parts knowledge, Toolbox DCCS, Toolbox Flanker, CRT-D, denominator neglect, base-rate sensitivity, actively open-minded thinking, living-things knowledge, backward digit span, and Need for Cognition. Children in the second grade and above completed the mathematical-equivalence problems between the living-things and backward-digit-span tasks.

Results

Descriptive statistics

Forty-one of 152 children provided incomplete data, resulting in missing values for 4.7% of the entire data set (not including actively open-minded thinking and the vitalist-biology composite). Twenty-four children were missing scores from one (18 children) or both (six children) NIH Toolbox tasks because of equipment failure or experimenter error. Eleven children were missing values from one (six children) or both (five children) of the Body Parts Interview and verbal fluency task because of audio-recording failures. Three children did not want to attempt the mathematical-equivalence problems. Two parents did not provide age information for their children. Finally, nine children did not complete the session because of fatigue or parent interruption and thus were missing multiple measures. Table 2 presents summary statistics of and bivariate Pearson correlations among our variables.

Children's CRT-D performance was as expected. For each item, the vast majority of children generated either the correct or the intuitive incorrect response (Table 1). Additionally, children at every age generated fewer other incorrect responses than intuitive incorrect responses for each item. There was a single exception—5-year-olds generated more other incorrect responses than intuitive incorrect responses for Question 3 (i.e., "A farmer has 5 sheep, all but 3 run away. How many are left?"). Thus, the CRT-D followed the fundamental structure of the original CRT for each age group. Further,

as an individual-differences measure, CRT-D scores demonstrated considerable variability within age (Fig. 1). Mean CRT-D scores increased by approximately 2.5 items from ages 5 to 12 years. With the exception of Questions 1 and 4, correct responding on individual CRT-D items was weakly to moderately correlated with age ($r_s = .18-.43$). Unlike adult CRT scores, CRT-D scores showed no gender difference (female: $M = 2.52$, male: $M = 2.48$).

CRT-D performance was moderately to strongly correlated with age, Toolbox DCCS, Toolbox Flanker, verbal fluency, backward digit span, denominator neglect, base-rate sensitivity, vitalist-biology composite, body-parts knowledge, living-things knowledge, and mathematical equivalence (i.e., all measures except for Need for Cognition; Table 2). In addition, every other predictor variable was also moderately to strongly correlated with one or more of the vitalist-biology and mathematical-equivalence outcomes (Table 2).

Regression analyses

We used Bayesian estimation to examine the predictive utility of children's cognitive reflection for conceptual understanding of vitalist biology and mathematical equivalence. Among other advantages, Bayesian analysis provided a common framework to take advantage of information from children with partial missing data (i.e., jointly modeling missing predictor data; McElreath, 2016) and evaluate predictive performance (e.g., regularizing priors, Bayesian leave-one-out [LOO] cross-validation, and projective predictive selection; see below). Full details of our Bayesian analyses can be found in the Supplemental Material available online. General results were replicated using classical frequentist methods (see the Supplemental Material). Data and R scripts to reproduce all analyses are available at <https://osf.io/e72ka/>.

We fitted Bayesian regression models for each outcome measure. All predictor variables, the vitalist-biology composite score, and the body-parts-knowledge score were scaled to have a mean of 0 and a standard deviation of 1. We used regularizing weakly informative priors for all regression parameters. Weakly informative priors contain enough information to rule out unreasonable parameter values but are not strong enough to rule out parameter values that might be relevant. For linear regression, we used $\text{normal}(\mu = 0, \sigma = 1)$ as the prior for beta coefficients. Thus, we expected that a 1-standard-deviation change in x would plausibly predict somewhere between a -2 -standard-deviation and $+2$ -standard-deviation change in outcome y (because roughly 95% of the prior distribution's probability mass lies between ± 2). For logistic regression, we used $\text{normal}(\mu = 0, \sigma = 2.5)$ as the prior for beta coefficients. This prior reflects disciplinary

Table 2. Summary Statistics and Bivariate Correlations for Study Variables

Variable	<i>n</i>	<i>M</i>	<i>SD</i>	Correlations														
				1	2	3	4	5	6	7	8	9	10	11	12			
1. GRT-D	152	2.50	1.75	—														
2. Age (months)	150	97.79	23.36	.50 [.37, .61]	—													
3. Toolbox DCCS	145	82.74	19.20	.36 [.21, .49]	.57 [.45, .67]	—												
4. Toolbox Flanker	129	84.02	15.40	.36 [.20, .50]	.62 [.51, .72]	.64 [.53, .73]	—											
5. Verbal fluency	144	13.75	5.59	.43 [.29, .56]	.54 [.41, .65]	.40 [.24, .53]	.39 [.23, .53]	—										
6. Backward digit span	143	3.51	1.29	.42 [.27, .54]	.55 [.42, .65]	.58 [.46, .68]	.51 [.37, .63]	.37 [.21, .51]	—									
7. Denominator neglect	148	3.59	2.19	.29 [.13, .43]	.40 [.26, .53]	.42 [.27, .55]	.37 [.21, .52]	.35 [.20, .49]	.39 [.25, .52]	—								
8. Base-rate sensitivity	147	2.44	1.39	.23 [.07, .38]	.21 [.05, .36]	.27 [.11, .42]	.25 [.07, .40]	.24 [.08, .39]	.25 [.09, .40]	.17 [.01, .33]	—							
9. Need for Cognition	143	3.13	0.56	.12 [−.05, .28]	.11 [−.06, .27]	.13 [−.04, .30]	.13 [−.05, .30]	.12 [−.05, .29]	.26 [.10, .40]	.11 [−.06, .27]	.10 [−.06, .26]	—						
10. Vitalist-biology composite	137	0.02	0.83	.51 [.37, .62]	.50 [.36, .61]	.48 [.34, .61]	.49 [.34, .62]	.48 [.33, .60]	.47 [.33, .59]	.24 [.07, .39]	.29 [.13, .44]	.11 [−.06, .27]	—					
11. Body-parts knowledge	144	4.55	1.77	.46 [.32, .58]	.55 [.42, .65]	.45 [.31, .58]	.43 [.27, .56]	.50 [.37, .62]	.42 [.28, .55]	.27 [.10, .41]	.26 [.10, .41]	.12 [−.05, .28]	.82 [.75, .87]	—				
12. Living-things knowledge	145	2.95	0.90	.38 [.23, .51]	.28 [.12, .42]	.30 [.14, .45]	.36 [.19, .50]	.26 [.09, .41]	.35 [.20, .49]	.12 [−.04, .28]	.19 [.03, .35]	.07 [−.10, .23]	.83 [.76, .87]	.35 [.19, .49]	—			
13. Mathematical equivalence	86	1.71	1.75	.38 [.18, .55]	.21 [−.00, .41]	.43 [.24, .59]	.40 [.19, .57]	.19 [−.03, .39]	.24 [.03, .43]	.27 [.06, .46]	.18 [−.04, .37]	.26 [.05, .45]	.36 [.16, .54]	.19 [−.03, .39]	.39 [.19, .56]	—		

Note: Values in brackets are 95% confidence intervals. GRT-D = Cognitive Reflection Test-Developmental; DCCS = Dimensional Change Card Sort.

knowledge that logistic regression coefficients are almost always between -5 and 5 on the logit scale (i.e., from probabilities .01 to .99).

For each outcome measure, we fitted a full model that included all predictor variables (i.e., CRT-D, age, Toolbox DCCS, Toolbox Flanker, verbal fluency, backward digit span, denominator neglect, base-rate sensitivity, and Need for Cognition). We report 95% credible intervals and median posterior point estimates for the coefficients of the full models. For a given model and data, we can be 95% certain that the true value of a parameter is contained within its 95% credible interval.

We additionally fitted control models that included all predictor variables besides CRT-D. We compared the predictive performance of the full and control models using an efficient Bayesian approximation of LOO cross-validation (Vehtari, Gelman, & Gabry, 2017). LOO cross-validation gives an (almost) unbiased estimate of predictive error for each model and, furthermore, allows us to compute the difference in predictive performance (with standard error) between two models. We report differences between models and their standard errors on the deviance scale (called *leave-one-out information criterion* [LOOIC]). Analogous to other difference statistics, a difference of less than 1 standard error suggests that the models have roughly similar predictive performance, whereas a difference of more than 2 standard errors suggests that one model is expected to have better predictive performance than the other.

Finally, we used projective predictive selection to identify and fit submodels that demonstrated similar (or better) out-of-sample predictive performance than the full reference models (Piironen & Vehtari, 2017). This method uses posterior information from a reference (full) model to find smaller candidate models with predictive distributions as close to the reference predictive distribution as possible (in terms of Kullback-Leibler divergence). The method first performs a forward search through the model space, starting from an empty model and at each step adding the variable that minimizes the predictive discrepancy between the submodel and reference model (conditional on previously entered variables). Next, LOO cross-validation and a decision criterion are used to select the final size of the submodel. Here, we first looked for the smallest submodel with better out-of-sample predictive performance than the reference model (i.e., ΔLOOIC more than 2 SE s above 0). If there was no better performing model, we selected the smallest submodel with predictive performance similar to that of the reference model (i.e., ΔLOOIC within 1 SE of 0). Overall, a submodel identified via projective predictive selection can give us an idea of how important the CRT-D is as a predictor relative to the other measured variables.

Vitalist-biology composite

We fitted linear regression models for children's biology-composite scores. Figure 2a displays the estimates of the full model. There was a positive effect of CRT-D: A 1-standard-deviation increase in CRT-D predicted a 0.28-standard-deviation increase in biology composite score, 95% credible interval = [0.12, 0.43]. CRT-D moderately improved the out-of-sample predictive performance of the full model in comparison with the control model ($\Delta\text{LOOIC} = 10.59$, $SE = 6.98$). Projective predictive selection suggested a submodel with age, CRT-D, Toolbox DCCS, verbal fluency, Toolbox Flanker, and base-rate sensitivity, in order of their relative predictive importance. The submodel provided moderately better predictive performance than the full model ($\Delta\text{LOOIC} = 3.96$, $SE = 2.56$). These results indicate that children's CRT-D performance had a positive effect on vitalist-biology understanding over and above the other measured variables. Further, conditional on children's age, CRT-D performance is likely the best predictor of overall vitalist-biology understanding among the other measured variables. We now look at the CRT-D's contribution to the subcomponents of vitalist biology—body-parts knowledge and living-things knowledge—individually.

Body-parts knowledge

We fitted linear regression models for children's body-parts knowledge. Figure 2b displays the estimates of the full model. There was a positive effect of CRT-D: A 1-standard-deviation increase in CRT-D predicted a 0.21-standard-deviation increase in body-parts knowledge, 95% credible interval = [0.05, 0.36]. CRT-D slightly improved the out-of-sample predictive performance of the full model in comparison with the control model ($\Delta\text{LOOIC} = 5.14$, $SE = 5.59$). Projective predictive selection suggested a submodel with age, verbal fluency, CRT-D, and Toolbox DCCS (ordered by relative predictive importance). The submodel provided better predictive performance than the full model ($\Delta\text{LOOIC} = 9.04$, $SE = 2.49$). These results indicate that children's CRT-D performance had a positive effect on body-parts knowledge over and above the other measured variables and, further, that CRT-D performance is among the minimal set of most relevant predictors.

Living-things knowledge

For children's living-things knowledge, we fitted logistic multilevel models with varying intercepts for participant and category (i.e., animals, plants, natural phenomena, artifacts). Figure 2c displays the estimates of the full model. There was a positive effect of CRT-D: A 1-standard-deviation increase in CRT-D predicted a 0.43-log-odds

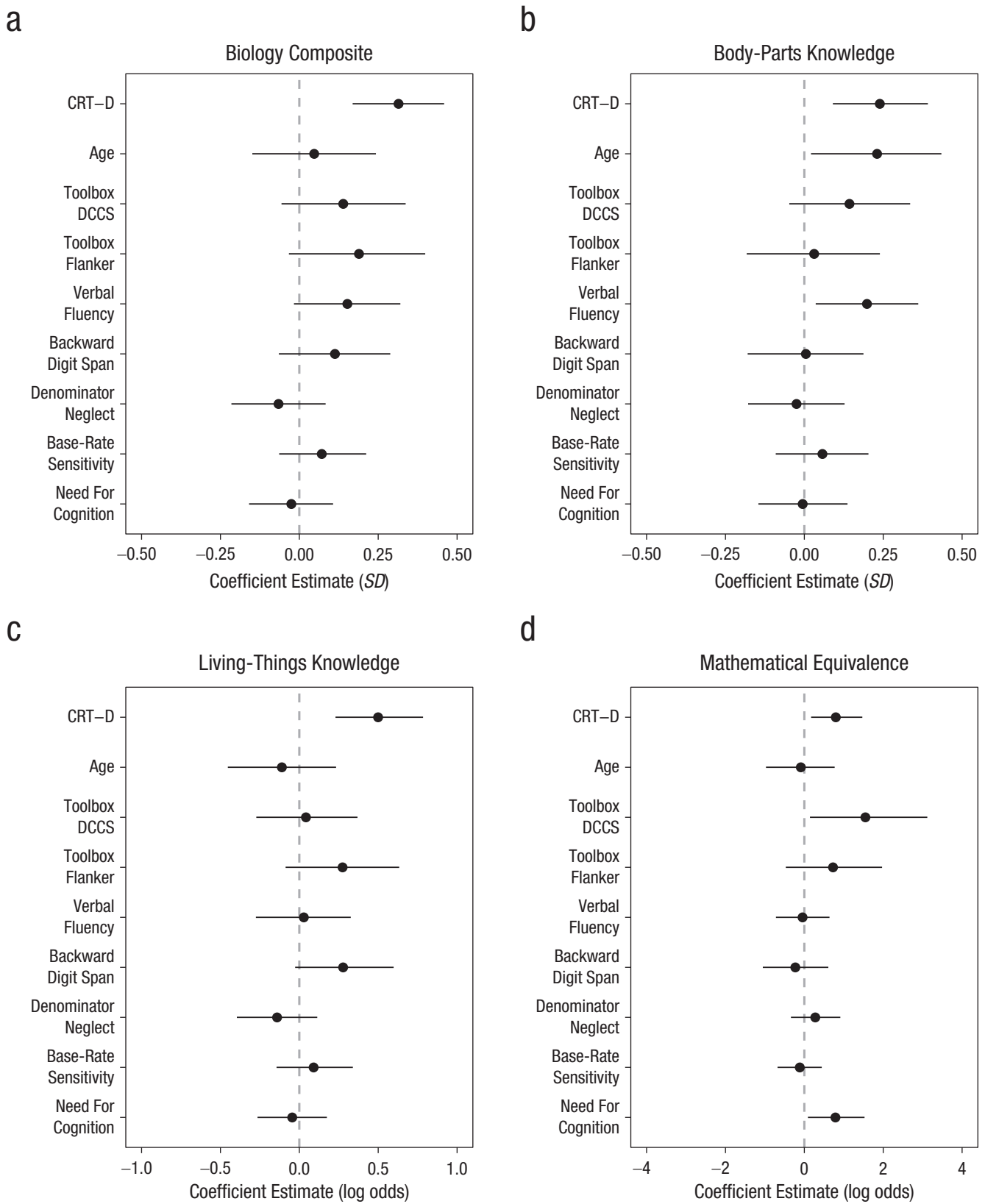


Fig. 2. Coefficient estimate (posterior median) for each variable included in the full model for (a) the biology composite, (b) body-parts knowledge, (c) living-things knowledge, and (d) mathematical equivalence. Error bars show 95% credible intervals. CRT-D = Cognitive Reflection Test-Developmental; DCCS = Dimensional Change Card Sort.

Table 3. Model Comparisons for Each Outcome Measure

Outcome	Control model	Full model	Submodel	Selected variables
	LOOIC	LOOIC	LOOIC	
Vitalist-biology composite	341.32 (19.75)	330.75 (18.48)	326.78 (18.28)	Age, CRT-D, Toolbox DCCS, verbal fluency, Toolbox Flanker, base-rate sensitivity
Body-parts knowledge	360.56 (15.74)	355.42 (15.43)	346.38 (15.76)	Age, verbal fluency, CRT-D, Toolbox DCCS
Living-things knowledge	585.61 (25.41)	578.58 (25.63)	568.49 (24.90)	CRT-D, Toolbox Flanker, backward digit span
Mathematical equivalence	116.77 (10.82)	113.32 (11.21)	101.45 (9.46)	Toolbox DCCS, CRT-D, Need for Cognition

Note: Standard errors are given in parentheses. A lower leave-one-out information criterion (LOOIC) reflects better out-of-sample prediction. Submodel variables are listed by order of selection entry (i.e., conditional improvement of out-of-sample prediction). CRT-D = Cognitive Reflection Test-Developmental; DCCS = Dimensional Change Card Sort.

increase in correct living-things knowledge, 95% credible interval = [0.14, 0.72]. CRT-D moderately improved the out-of-sample predictive performance of the full model in comparison with the control model ($\Delta\text{LOOIC} = 7.04$, $SE = 6.01$). Projective predictive selection suggested a submodel with CRT-D, Toolbox Flanker, and backward digit span (ordered by relative predictive importance). The submodel provided better predictive performance than the full model ($\Delta\text{LOOIC} = 10.09$, $SE = 3.52$). These results indicate that children's CRT-D performance had a positive effect on living-things knowledge over and above the other measured variables and, further, that CRT-D performance is likely the single best predictor of performance on the living-things judgment task among all other measures.

Mathematical equivalence

We fitted logistic regression models for children's mathematical-equivalence performance. We used generation of at least one correct strategy as the outcome because a major shift in conceptual understanding of the equal sign is required to solve just one problem correctly (McNeil et al., 2019). Descriptively, 45% of children solved all four problems incorrectly, 27% solved one to three problems correctly, and 28% solved all problems correctly. Figure 2d displays the estimates of the full model. There was a positive effect of CRT-D: A 1-standard-deviation increase in CRT-D predicted a 0.83-log-odds increase in generating a correct strategy, 95% credible interval = [0.18, 1.51]. CRT-D slightly improved the out-of-sample predictive performance of the full model in comparison with the control model ($\Delta\text{LOOIC} = 3.46$, $SE = 6.21$). Projective predictive selection suggested a submodel with Toolbox DCCS, CRT-D, and Need for Cognition (ordered by relative predictive importance). The submodel provided better predictive performance than the full model ($\Delta\text{LOOIC} = 11.87$, $SE = 4.06$). These results indicate that children's CRT-D performance had a positive effect on mathematical-equivalence performance

over and above the other measured variables and, further, that CRT-D performance is among the minimal set of most relevant predictors.

Table 3 provides a summary of model comparisons.

Discussion

In this research, we examined whether a recent measure of school-age children's cognitive reflection, the CRT-D, predicts conceptual understanding in science and mathematics. To do so, we measured children's CRT-D performance, executive functions, rational thinking, and conceptual understanding of vitalist biology and mathematical equivalence, critical domains of early science and mathematics in which conceptual change is protracted and hard won. We found that CRT-D performance was a strong predictor of children's conceptual understanding in both domains, even when adjusting for age, executive functions, and rational thinking. Further, in both domains, the CRT-D was among the most important variables for out-of-sample prediction. Our findings suggest that the CRT-D successfully measures children's cognitive reflection and that cognitive reflection is a valuable construct in the study of conceptual development.

Our results are consistent with two explanations of how cognitive reflection supports children's conceptual understanding. First, cognitive reflection may facilitate children's expression of counterintuitive concepts, as executive-function skills have been shown to do (Vosniadou et al., 2018). The ability to reflect on and override an intuitive response almost certainly supports children's biological reasoning (in which animism conflicts with vitalism) and mathematical-equivalence problem solving (in which operational conceptions of the equal sign conflict with relational ones).

A second possibility is that cognitive reflection facilitates children's initial learning of counterintuitive scientific and mathematical ideas. Children with greater cognitive reflection may respond to counterintuitive

experience or instruction differently from less reflective children. Previous research demonstrates that executive-function skills are involved in both the expression and the construction of children's science and math concepts (Bascandzjev et al., 2018; Cragg & Gilmore, 2014), and we suspect that cognitive reflection plays both roles as well. Indeed, we have found that cognitive reflection predicts children's learning of counterintuitive scientific concepts, even after adjusting for baseline conceptual knowledge (Young & Shtulman, 2020).

One potential limitation to our study is that certain CRT-D items require factual knowledge relevant to our conceptual measures. For example, Question 5 ("What do cows drink?") requires factual biological knowledge potentially relevant to vitalist biology, and Question 3 ("A farmer has 5 sheep, all but 3 run away. How many are left?") requires mathematical knowledge potentially relevant to mathematical equivalence. It is possible that the CRT-D's predictive power was driven by a few domain-relevant items and not the CRT-D more broadly. Reanalysis of the vitalist-biology outcomes using a modified CRT-D without the biology-relevant Questions 5 and 6 and the mathematical-equivalence outcome using a modified CRT-D without the math-relevant Questions 3 and 4 yielded similar results (see the Supplemental Material). Thus, the predictive utility of the CRT-D was not restricted to domain-relevant items. However, future research would benefit from measuring and controlling for children's factual knowledge of the relevant domains.

Performance on the adult CRT is correlated with executive-function skills; however, the predictive power of the CRT is often separable from these abilities (e.g., Toplak et al., 2011). The present findings extend this pattern, showing that the predictive utility of the CRT-D extends beyond executive function and rational thinking, at least as operationalized in the current study. What may account for the CRT-D's surprisingly unique predictive power?

Ongoing research with adults is focused on specifying the processes and components that underlie cognitive reflection. Whether cognitive reflection is a set of cognitive abilities, dispositions, acquired knowledge, or some combination thereof is a matter of controversy and active debate. Popular accounts of what the CRT captures include (a) cognitive miserliness, or an unwillingness to go beyond heuristic processing and invest cognitive effort (Stanovich et al., 2016); (b) a general disposition toward analytic thinking (Pennycook et al., 2015); and (c) ignorance of the relevant rules, strategies, or beliefs that facilitate correct responding (Szasz et al., 2017). Recent research has also revealed that some individuals never consider an intuitive response to CRT items (Bago & De Neys, 2019; Szasz et al., 2017). Such intuitively logical individuals generate a correct

response under cognitive load and time pressure but generate correct justifications only after deliberation (Bago & De Neys, 2019). Thus, cognitive reflection may be used to override an inaccurate intuitive response or to look for an explicit justification supporting a correct response (De Neys & Pennycook, 2019). The present study was not designed to decide among these accounts, but programmatic research on the development of cognitive reflection will help adjudicate debates over its underlying mechanisms.

Broadly, this study highlights the value and potential of extending the study of cognitive reflection from adults to children. For adults, the CRT is a unique predictor of a broad range of outcomes and behaviors. If children's cognitive reflection is continuous with that of adults, the CRT-D may prove useful in examining the development of domains as far-ranging as stereotyping (Hammond & Cimpian, 2017), moral reasoning (Royzman et al., 2015), and evidence evaluation (Pennycook & Rand, 2019). Children's cognitive reflection may also prove to be an effective target for intervention.

That said, the CRT-D used in the present research should not be viewed as a fixed scale but one open to revision. We took a conservative approach and retained all CRT-D items because they matched the response structure of the original CRT across our age range. The observed discrepancy between a relatively modest Cronbach's α and satisfactory McDonald's ω_{total} suggests that the CRT-D, like the original CRT, may be psychometrically complex (Stanovich, 2018; Stuppel et al., 2017). Still, additional research is needed to verify that our CRT-D items function as intended across the targeted age range. For example, only about 8% of children in the present study responded that a pound of rocks weighs the same as a pound of feathers (Question 6). Younger children may have lacked the knowledge required to answer correctly, whereas older children may have failed to detect and override the intuitive response. A related concern is that certain items in the present CRT-D might not function well in other cultures or languages. Various methods have been used to address such issues with adult CRTs, including chronometric analysis (Stuppel et al., 2017; Travers et al., 2016), protocol analysis (Szasz et al., 2017), and item-response theory (Primi et al., 2016). Future research with larger and more diverse samples should apply such methods to better understand the psychometric properties of the CRT-D and guide revision of the scale.

Conclusion

The CRT is the predominant measure of adult individual differences in analytic versus intuitive thinking. Here, we examined a CRT for school-age children, the CRT-D,

and found that it was a strong and unique predictor of children's conceptual understanding in science and mathematics. These findings demonstrate, for the first time, the theoretical and practical importance of children's cognitive reflection. We anticipate that the CRT-D will allow future researchers to investigate not only individual differences in cognitive abilities among children but also the development, malleability, and consequences of cognitive reflection more generally.

Transparency

Action Editor: Erika E. Forbes

Editor: D. Stephen Lindsay

Author Contributions

Both authors designed the research. A. G. Young oversaw data collection and analyzed the data. A. G. Young drafted the manuscript, and A. Shtulman provided critical revisions. Both authors approved the final manuscript for submission.

Declaration of Conflicting Interests

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

Funding


The James S. McDonnell Foundation supported this research through an Understanding Human Cognition Scholar Award to A. Shtulman.

Open Practices

All data and analysis code have been made publicly available via OSF and can be accessed at <https://osf.io/e72ka>. The design and analysis plans for this study were not preregistered. This article has received the badges for Open Data and Open Materials. More information about the Open Practices badges can be found at <http://www.psychologicalscience.org/publications/badges>.



ORCID iD

Andrew G. Young  <https://orcid.org/0000-0003-1376-8839>

Acknowledgments

We thank the families who participated in this research and the Kidspace Children's Museum. We also thank Dallas Boyce, Katherine Clark, Julia Moreland, Lesley Pilgrim, Allison Powers, and Beatrice Terino for their assistance with data collection.

Supplemental Material

Additional supporting information can be found at <http://journals.sagepub.com/doi/suppl/10.1177/0956797620954449>

References

- Alloway, T. P., Gathercole, S. E., Kirkwood, H., & Elliott, J. (2009). The cognitive and behavioral characteristics of

- children with low working memory. *Child Development*, *80*, 606–621. doi:10.1111/j.1467-8624.2009.01282.x
- Bago, B., & De Neys, W. (2019). The Smart System 1: Evidence for the intuitive nature of correct responding on the bat-and-ball problem. *Thinking & Reasoning*, *25*, 257–299.
- Bascandziev, I., Tardiff, N., Zaitchik, D., & Carey, S. (2018). The role of domain-general cognitive resources in children's construction of a vitalist theory of biology. *Cognitive Psychology*, *104*, 1–28. doi:10.1016/j.cogpsych.2018.03.002
- Corgnet, B., Espín, A. M., & Hernán-González, R. (2015). The cognitive basis of social behavior: Cognitive reflection overrides antisocial but not always prosocial motives. *Frontiers in Behavioral Neuroscience*, *9*, Article 287. doi:10.3389/fnbeh.2015.00287
- Cragg, L., & Gilmore, C. (2014). Skills underlying mathematics: The role of executive function in the development of mathematics proficiency. *Trends in Neuroscience and Education*, *3*, 63–68. doi:10.1016/j.tine.2013.12.001
- De Neys, W., & Pennycook, G. (2019). Logic, fast and slow: Advances in dual-process theorizing. *Current Directions in Psychological Science*, *28*, 503–509.
- Don, H. J., Goldwater, M. B., Otto, A. R., & Livesey, E. J. (2016). Rule abstraction, model-based choice, and cognitive reflection. *Psychonomic Bulletin & Review*, *23*, 1615–1623. doi:10.3758/s13423-016-1012-y
- Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. doi:10.1257/089533005775196732
- Gervais, W. M. (2015). Override the controversy: Analytic thinking predicts endorsement of evolution. *Cognition*, *142*, 312–321. doi:10.1016/j.cognition.2015.05.011
- Gómez-Chacón, I. M., García-Madruga, J. A., Vila, J. Ó., Elosúa, M. R., & Rodríguez, R. (2014). The dual processes hypothesis in mathematics performance: Beliefs, cognitive reflection, working memory and reasoning. *Learning and Individual Differences*, *29*, 67–73.
- Hammond, M. D., & Cimpian, A. (2017). Investigating the cognitive structure of stereotypes: Generic beliefs about groups predict social judgments better than statistical beliefs. *Journal of Experimental Psychology: General*, *146*, 607–614. doi:10.1037/xge0000297
- Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, *8*, 188–201.
- Hornburg, C. B., Wang, L., & McNeil, N. (2018). Comparing meta-analysis and individual person data analysis using raw data on children's understanding of equivalence. *Child Development*, *89*, 1983–1995. doi:10.1111/cdev.13058
- Inagaki, K., & Hatano, G. (2002). *Young children's naive thinking about the biological world*. New York, NY: Psychology Press.
- Keller, U., Strobel, A., Wollschläger, R., Greiff, S., Martin, R., Vainikainen, M.-P., & Preckel, F. (2019). A Need for Cognition scale for children and adolescents: Structural analysis and measurement invariance. *European Journal of Psychological Assessment*, *35*, 137–149. doi:10.1027/1015-5759/a000370

- Knuth, E. J., Stephens, A. C., McNeil, N. M., & Alibali, M. W. (2006). Does understanding the equal sign matter? Evidence from solving equations. *Journal for Research in Mathematics Education*, *37*, 297–312.
- McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan*. Boca Raton, FL: Chapman & Hall/CRC Press.
- McNeil, N. M., & Alibali, M. W. (2005). Why won't you change your mind? Knowledge of operational patterns hinders learning and performance on equations. *Child Development*, *76*, 883–899. doi:10.1111/j.1467-8624.2005.00884.x
- McNeil, N. M., Hornburg, C. B., Devlin, B. L., Carrazza, C., & McKeever, M. O. (2019). Consequences of individual differences in children's formal understanding of mathematical equivalence. *Child Development*, *90*, 940–956. doi:10.1111/cdev.12948
- Munakata, Y., Snyder, H. R., & Chatham, C. H. (2012). Developing cognitive control. *Current Directions in Psychological Science*, *21*, 71–77. doi:10.1177/0963721412436807
- Pennycook, G., Fugelsang, J. A., & Koehler, D. J. (2015). Everyday consequences of analytic thinking. *Current Directions in Psychological Science*, *24*, 425–432. doi:10.1177/0963721415604610
- Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition*, *188*, 39–50. doi:10.1016/j.cognition.2018.06.011
- Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*, 711–735. doi:10.1007/s11222-016-9649-y
- Primi, C., Morsanyi, K., Chiesi, F., Donati, M. A., & Hamilton, J. (2016). The development and testing of a new version of the Cognitive Reflection Test applying item response theory (IRT). *Journal of Behavioral Decision Making*, *29*, 453–469. doi:10.1002/bdm.1883
- Royzman, E. B., Landy, J. F., & Leeman, R. F. (2015). Are thoughtful people more utilitarian? CRT as a unique predictor of moral minimalism in the dilemmatic context. *Cognitive Science*, *39*, 325–352. doi:10.1111/cogs.12136
- Shtulman, A., & McCallum, K. (2014). Cognitive reflection predicts science understanding. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 2937–2942). Austin, TX: Cognitive Science Society.
- Sinatra, G. M., Southerland, S. A., McConaughy, F., & Demastes, J. W. (2003). Intentions and beliefs in students' understanding and acceptance of biological evolution. *Journal of Research in Science Teaching*, *40*, 510–528. doi:10.1002/tea.10087
- Slaughter, V., & Lyons, M. (2003). Learning about life and death in early childhood. *Cognitive Psychology*, *46*, 1–30. doi:10.1016/S0010-0285(02)00504-2
- Stanovich, K. E. (2018). Miserliness in human cognition: The interaction of detection, override and mindware. *Thinking & Reasoning*, *24*, 423–444.
- Stanovich, K. E., West, R. F., & Toplak, M. E. (2016). *The rationality quotient: Toward a test of rational thinking*. Cambridge, MA: MIT Press.
- Stuppel, E. J., Pitchford, M., Ball, L. J., Hunt, T. E., & Steel, R. (2017). Slower is not always better: Response-time evidence clarifies the limited role of miserly information processing in the Cognitive Reflection Test. *PLOS ONE*, *12*(11), Article e0186404. doi:10.1371/journal.pone.0186404
- Szaszi, B., Szollosi, A., Palfi, B., & Aczel, B. (2017). The Cognitive Reflection Test revisited: Exploring the ways individuals solve the test. *Thinking & Reasoning*, *23*, 207–234.
- Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the Cognitive Reflection Test. *Judgment and Decision Making*, *11*, 99–113.
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. *Memory & Cognition*, *39*, 1275–1289. doi:10.3758/s13421-011-0104-1
- Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Rational thinking and cognitive sophistication: Development, cognitive abilities, and thinking dispositions. *Developmental Psychology*, *50*, 1037–1048. doi:10.1037/a0034910
- Travers, E., Rolison, J. J., & Feeney, A. (2016). The time course of conflict on the Cognitive Reflection Test. *Cognition*, *150*, 109–118. doi:10.1016/j.cognition.2016.01.015
- Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, *27*, 1413–1432. doi:10.1007/s11222-016-9696-4
- Vosniadou, S., Pnevmatikos, D., Makris, N., Lepenioti, D., Eikospentaki, K., Chountala, A., & Kyrianiakis, G. (2018). The recruitment of shifting and inhibition in on-line science and mathematics tasks. *Cognitive Science*, *42*, 1860–1886. doi:10.1111/cogs.12624
- Wong, T. T.-Y. (2018). Is conditional reasoning related to mathematical problem solving? *Developmental Science*, *21*, Article e12644. doi:10.1111/desc.12644
- Young, A. G., Powers, A., Pilgrim, L., & Shtulman, A. (2018). Developing a Cognitive Reflection Test for school-age children. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1232–1237). Austin, TX: Cognitive Science Society.
- Young, A. G., & Shtulman, A. (2020). How children's cognitive reflection shapes their science understanding. *Frontiers in Psychology*, *11*, Article 1247. doi:10.3389/fpsyg.2020.01247
- Zaitchik, D., Iqbal, Y., & Carey, S. (2014). The effect of executive function on biological reasoning in young children: An individual differences study. *Child Development*, *85*, 160–175. doi:10.1111/cdev.12145
- Zelazo, P. D., Anderson, J. E., Richler, J., Wallner-Allen, K., Beaumont, J. L., & Weintraub, S. (2013). II. NIH Toolbox Cognition Battery (CB): Measuring executive function and attention. *Monographs of the Society for Research in Child Development*, *78*, 16–33. doi:10.1111/mono.12032
- Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α , Revelle's β , McDonald's ω_i : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*, 123–133.