CoolThink@JC Pilot Evaluation

Endline Report Appendices

May 2020

Contents

APPENDIX A: *COOLTHINK@JC* FRAMEWORK

CT Concepts **The concepts designers engage with as they program.**	
1. Sequences	Identify a series of ordered steps for solving a programming task
2. Events	One thing causing another thing to happen
3. Conditionals	Make decision based on conditions
4. Operators	Support for mathematical and logical expressions
5. Parallelism	Make things happen at the same time
6. Repetition	Run the same sequence multiple times
7. Naming and Variables	Name objects, procedures, or values, use variables
8. Data Structures	Basic ways data are stored, retrieved, and updated
9. Procedures	Create code blocks to modularize and abstract sequences of commands

CT Practices: Problem-solving Skills **The practices designers develop as they engage with the concepts.**	
1. Testing and Debugging	Make sure things work, otherwise find and solve problems when they arise
2. Being incremental and Iterative	Develop a little bit, then try it out, then develop more
3. Reusing and Remixing	Make something by building on existing code, projects or ideas
4. Abstracting and Modularizing	Explore connections between the whole and the parts
5. Algorithmic Thinking	Articulate a problem's solution in well-defined rules and steps

CT Perspectives: Identity and Motivation **The perspectives designers form about the world around them and about themselves.**	
1. Expressing	Create and express idea through this new medium
2. Questioning	Feel empowered to ask questions about and with technology
3. Connecting	Appreciate that others are engaged with and appreciate one's creations
4. Digital Empowerment	Develop the ability to see problems in the world as solvable through coding
5. Computational Identity	See oneself as being able to enhance the world through coding

APPENDIX B: ANALYTICAL METHODS

This appendix introduces several of the analytical methods used in this research.

Partial Matrix Sampling

A matrix sampling approach to computational thinking (CT) assessment design involves distributing sets of items across multiple forms and randomly assigning the forms to students. This process allows data to be collected on more items than can be administered to one student at a time. This approach is commonly used when the goal is to determine how cohorts of students are performing rather than to obtain individual scores for students. For example, this approach is used in the U.S. National Assessment for Educational Progress, an assessment that measures student achievement across the United States.

While this method does not allow for direct comparison of students, as different students will receive different items which may cover a different set of concepts, it is well tuned for comparisons of the performance of groups of students. A version of matrix sampling, referred to as *partial matrix sampling*, allows for some individual student comparison. In in this version there is a set of common items, or items that all students receive, and the rest of the items are split up across forms. Using this approach along with an item response theory analysis (described below) allows student scores to be generated that are comparable across students even if they do not take all of the same items. Specifically, student ability is estimated using item response models based the items they take. The items that they have in common are used to anchor their ability estimates so that the estimates are on the same scale and thus comparable. We used the partial matrix sampling approach to develop the CT Concepts and CT Practices assessments. The benefit of this approach is that it allows for measurement of the cohort of students on all items while reducing the testing time for individual students.

Over the 3 years of this evaluation, the number of assessment items was reduced to keep pace with ongoing streamlining of the concepts covered in the curriculum. For the CT Concepts assessment, beginning in 2018 the topic of Procedures was dropped, which in turn made it possible for students to complete the full set of CT Concepts items in one testing period. As a result, in 2018 and 2019 partial matrix sampling was only needed for CT Practices.

Item Response Theory

The students' CT Concepts, Practices, and Perspectives scores were calculated based on Item Response Theory (IRT). IRT is a latent variable modeling approach by which scores on assessment items are used to place items on a scale indicating their difficulty, as well as to place students on the same scale indicating their ability. This method of analysis allows us to create an overall measure of computational thinking ability, and to look at the progression of an individual student or cohort of students along that continuum. IRT is tuned to handle missing data, which is important in matrix sampling because individual students will only respond to a subset of the total pool of test items or constructs on a given assessment. Student ability is estimated based on the student's available responses. The responses that are missing by design will not contribute to ability estimation. This allows us to generate an overall estimate of computational thinking ability for each student at each administration point.

With this design we can compare individual students' progress along the full continuum of computational thinking ability. Because matrix sampling randomly distributes items that measure individual constructs across a large number of students, we can also compare cohorts of students on each construct. It is important to recognize that this is different than designs that track individual students' learning of each specific construct over time.

In order to maintain comparability in estimates of item difficulty and student abilities across different assessments, we include common items (often referred to in IRT as anchor items) that help define the scale. Therefore, for the CT Concepts assessments we not only have common items across forms for the assessments administered at the same level, we also include items that are the same across levels. Using these anchor items, we are able to calibrate item characteristics such as difficulty and discrimination with different student samples, and to link assessments across forms and years. In addition, we calculated the estimated values for each analysis using all of the data (across years and forms) together. This ensures that the item difficulties, which are used to calculate student ability estimates, are comparable. Item difficulties for items that overlap are the same across forms and years, and those item difficulties are used in part to generate item difficulties for items that appear only on one form. We can then see the variation in the ability estimates of the students based on one common continuum of the construct of interest.

For CT Concepts with dichotomous items, we used a one-parameter logistic (1PL) model as we were concerned with the difficulty of the items. For CT Practices, items are either dichotomously scored or polytomously scored up to 5 points. We used a partial credit model (PCM) to account for the items with multiple scoring categories. For CT Perspectives where items were on a Likert-type scale, we used a rating scale model to account for the multiple response categories of the items. Additional details on these models and their uses can be found in Embretson & Reise (2013).[1]

Conversion to NCE Scores

To make the IRT scores more interpretable, we converted them into Normal Curve Equivalent (NCE) scores. An NCE score ranges from 0–100 and follows a normal distribution, with a mean of 50 and a standard deviation of 21.06. Converting the IRT scores to the NCE scale puts the scores in an easy-to-interpret form and facilitates the comparison of the magnitude of findings among different constructs.

CT Perspectives is measured in seven subscales: students' interest in programming, digital self-efficacy, utility motivation, motivation to help the world, creativity, engagement, and belonging. To create an NCE score that summarized students' overall positive perspectives about computational thinking, we first standardized the IRT scores for each subconstruct, then averaged the standardized scores for a combined CT Perspectives score, and finally converted the combined CT Perspectives score into the NCE score. The composite measure of CT Perspectives reported here includes the first six subconstructs listed above. It does not include the Belonging construct. Rather than focusing on students' opinions of programming, the Belonging subscale relates to students' opinions of working with others while they are programming. For this reason, Belonging showed low correlation with the other subscales and was eliminated from the overall measure.

[1] Embretson, S. E., & Reise, S. P. (2013). *Item response theory for psychologists.* London: Psychology Press.

Hierarchical Linear Models for Impact Analysis

The Year 2 and Year 3 impact estimates were derived from a two-level hierarchical model with student and school levels that controlled for covariates at their own levels. The model is shown below:

where *i* is students, *j* is schools; Y_{ij} is a student outcome; $CoolThink_{ij}$ equals 1 for *CoolThink@JC* pilot schools and 0 for comparison schools; e_{ij} and r_j are student and school random effects. β_*k* represents a vector of student covariates; β1 is the estimated impact of *CoolThink@JC* on the student outcome.

Exhibit B1 lists the control variables we included in the hierarchical model at each level.

$$y_{ij} = \beta_0 + \beta_1(CoolThink_{ij}) + \beta_k(k^{th} student\ covariate_{ij}) + \beta_l(l^{th} school\ covariate_j) + e_{ij} + r_j$$

Exhibit B1. Control Variables for the Impact Model

Level	Control Variables
School	% of students using financial aid % of students with special needs % of non-Chinese speakers School prior coding instruction experience Paper vetting score Cohort indicator in combined cohorts analysis
Student	Baseline outcome measure Gender Grade level

APPENDIX C: MODEL ESTIMATES

This appendix shows the baseline equivalence, model estimates, standard errors and p-values for the impact analysis.

CT Concepts

We examined the difference in student outcome scores between *CoolThink@JC* pilot and comparison schools at baseline to check whether the two groups were equivalent before the start of the intervention. Exhibit C1 shows the baseline Concepts scores and sample sizes for pilot and comparison schools respectively. The difference between the two groups is within 0.25 standard deviation, so this analysis achieved baseline equivalence with adjustment of the baseline Concepts score.

Exhibit C2 shows the impact and subgroup differential impact estimates from a two-level HLM model with student and school levels. This analysis combines pilot schools in both Cohort 1 and Cohort 2. These results were discussed in the main text.

Exhibit C1. Baseline CT Concepts Scores, Year 2 Impact Analysis

	Mean	SD	Students	Schools
Pilot	38.8	12.9	4776	30
Comparison	37.1	12.1	3517	22

Exhibit C2. Model Estimates for CT Concepts Scores, Year 2 Impact Analysis

	Estimate	SE	P-value
Overall impact			
Pilot	5.0	2.9	0.084
Differential impact on subgroups			
Girls versus boys	-2.3*	1.0	0.020
Grade 5 versus grade 6	1.6	1.0	0.108
Baseline concepts score	0.0	0.0	0.209
Standardized baseline math test score	0.7	0.5	0.169
School % students with special needs	0.5	0.3	0.144
School % students using financial aid	-0.1	0.1	0.220
Ever done programming	-0.3	1.1	0.791
Ever done programming in class	-0.4	1.2	0.752
Have internet at home	1.5	1.2	0.221
Computer use at home	0.2	0.3	0.448
School existing coding experience	-0.6	0.6	0.304

** p < 0.05.*

For the subgroup analyses, please note that school % non-Chinese speaking students is not included because the vast majority of schools reported no, or fewer than 5%, non-Chinese speaking students. This does not provide sufficient variation for analysis.

Exhibit C3 shows the baseline Concepts scores and sample sizes for pilot and comparison schools respectively for Cohort 1 Year 3 analysis. The difference between the two groups are within 0.25 standard deviation, therefore this analysis achieved baseline equivalence with adjustment of the baseline Concepts score.

Exhibit C4 shows the impact estimates for Cohort 1 Year 3 analysis from a two-level HLM model with student and school levels. This result is consistent with that of Year 2 analysis combining Cohort 1 and Cohort 2.

Exhibit C3. Baseline CT Concepts Scores, Cohort 1 Year 3 Impact Analysis

	Mean	SD	Students	Schools
Pilot	38.6	12.0	764	9
Comparison	36.6	11.2	535	12

Exhibit C4. Model Estimates for CT Concepts Scores, Cohort 1 Year 3 Impact Analysis

	Estimate	SE	P-value
Overall impact			
Pilot	3.8	4.6	0.412

CT Practices

For CT practices, we only included Cohort 2 schools because students were first assessed in September 2018, when Cohort 1 pilot schools had already completed level 1. We therefore do not have baseline measure in CT Practices for Cohort 1 pilot schools.

Exhibit C5 shows baseline CT Practices scores and sample sizes for pilot and comparison schools respectively. The difference between the two groups is within 0.25 standard deviation, so this analysis achieved baseline equivalence with adjustment of the baseline CT Practices score.

Exhibit C6 shows the impact and subgroup differential impact estimates from a two-level HLM model with student and school levels. This result was discussed in the main text.

We also explored the impact of CoolThink on CT Practices after 3 years of intervention for Cohort 1 schools. Because there was no baseline CT Practices assessment for students in Cohort 1 schools, we conducted the analysis on the CT Practices outcome score while adjusting for baseline CT Concepts score. Exhibit C7 shows baseline practices scores and sample sizes for pilot and comparison schools respectively. The difference between the two groups is within 0.25 standard

Exhibit C5. Baseline CT Practices Scores, Cohort 2 Year 2 Impact Analysis

	Mean	SD	Students	Schools
Pilot	48.4	15.6	3900	20
Comparison	47.2	14.7	2487	20

Exhibit C6. Model Estimates for CT Practices Scores, Cohort 2 Year 2 Impact Analysis

	Estimate	SE	P-value
Overall impact			
Pilot	9.3***	2.8	0.001
Differential impact on subgroups			
Girls versus boys	0.6	1.2	0.605
Grade 5 versus grade 6	-1.3	1.2	0.283
Baseline Outcomes score	0.1*	0.0	0.041
Standardized baseline math test score	1.9**	0.4	0.002
School % students with special needs	0.2	0.4	0.533
School % students using financial aid	0.1	0.1	0.226
Ever done programming	-1.1	0.9	0.478
Ever done programming in class	-1.1	1.0	0.518
Have internet at home	0.9	0.9	0.570
Computer use at home	0.2	0.2	0.699
School existing coding experience	0.0	0.7	0.999

**p < 0.05, **p < 0.01, ***p < 0.001.*

Exhibit C7. Baseline Student Practices Scores, Cohort 1 Year 3 Impact Analysis

	Mean	SD	Students	Schools
Pilot	48.0	14.2	739	10
Comparison	45.7	13.3	665	12

deviation, so this analysis achieved baseline equivalence with adjustment of the baseline CT Practices score.

Exhibit C8 shows the impact estimates for Cohort 1 Year 3 analysis from a two-level HLM model with student and school levels. The estimated impact of 10.6 NCE points is similar in magnitude to the estimated impact of 9.3 NCE points for the Year 2 analysis with both cohorts of pilot schools, and is marginally significant with a p-value of 0.08.

Exhibit C8. Model Estimates for Student Practices Scores, Cohort 1 Year 3 Impact Analysis

	Estimate	SE	P-value
Overall impact			
Pilot	10.6	6.0	0.076

CT Perspectives

We examined the differences in student perspectives overall and sub-construct scores between *CoolThink@JC* pilot and comparison schools at baseline to see whether the two groups were equivalent in student outcomes before the start of the intervention. Exhibit C9 presents the descriptive statistics for pilot and comparison groups respectively. Across all perspective sub-constructs, the mean differences between the two groups are smaller than 0.25 of the respective standard deviations. We deemed the pilot and comparison schools achieved equivalence at baseline with adjustment of the baseline Perspectives scores.

Exhibit C9. Baseline CT Perspectives Scores, Year 2 Impact Analysis, 30 Pilot and 22 Comparison Schools

		Mean	SD	Students
Overall	Pilot	51.9	18.1	4914
	Comparison	49.0	17.7	3652
Belonging	Pilot	50.3	19.1	4114
	Comparison	49.4	17.8	2776
Interest	Pilot	51.2	21.0	4246
	Comparison	48.4	20.3	2922
Engagement	Pilot	52.3	21.1	4243
	Comparison	49.6	21.0	2922
Motivation World	Pilot	50.9	20.0	4249
	Comparison	48.6	19.7	2949
Efficacy	Pilot	50.1	20.7	4243
	Comparison	46.7	20.5	2944
Motivation Utility	Pilot	51.4	21.1	4234
	Comparison	49.2	20.9	2941

Exhibits C10 shows impact estimates for the overall CT Perspectives measure and its subconstructs from a two-level HLM model with student and school levels. The results were discussed in the main text.

Exhibit C10. Model Estimates for CT Perspectives Scores, Year 2

	Estimate	SE	P-value
Overall impact			
Perspectives	-2.3	1.5	0.131
Belonging	-2.2	1.6	0.175
Interest in Programming	-2.3	1.8	0.188
Engagement	-2.5	1.6	0.114
Motivation to help the world	-0.6	1.6	0.709
Creativity	-2.3	1.7	0.179
Self-efficacy*	-3.1	1.6	0.048
Utility motivation	-1.5	1.6	0.361

* *p < .05*

Exhibit C11 presents the descriptive statistics of perspectives overall and sub-constructs for pilot and comparison schools at baseline for Cohort 1 Year 3 analysis. For this cohort, pilot schools have substantially higher baseline scores (more than 0.25 standard deviation) than comparison schools on the overall perspectives scale and five of the seven sub-constructs (interest in programming, engagement, creativity, self-efficacy, and utility motivation).

The two exceptions to this lack of baseline equivalence are belonging and motivation to help the world. For these two sub-constructs, the differences between pilot and comparison schools are within 0.25 standard deviations with adjustment of the baseline CT Practices score. Due to the lack of baseline equivalence of the other CT perspective constructs, the results for Cohort 1 Year 3 analysis of CT Perspectives should be interpreted with caution.

Exhibits C12 shows the Cohort 1 Year 3 impact estimates on perspectives scales from a two-level HLM model with student and school levels. The results are consistent with those obtained in the Year 2 analysis.

Exhibit C11. Baseline CT Perspectives Scores, Cohort 1 Year 3 Impact Analysis, 10 Cohort 1 Pilot and 12 Comparison Schools

		Mean	SD	Students
Overall	Pilot	56.6	18.1	798
	Comparison	49.9	17.1	822
Belonging	Pilot	45.7	16.9	408
	Comparison	47.8	14.7	382
Interest	Pilot	61.0	22.5	410
	Comparison	52.0	21.2	398
Engagement	Pilot	59.6	21.1	410
	Comparison	52.3	21.6	399
Motivation World	Pilot	49.8	14.1	405
	Comparison	46.9	14.3	443
Efficacy	Pilot	50.4	15.3	405
	Comparison	46.5	14.3	442
Motivation Utility	Pilot	52.7	17.9	405
	Comparison	45.0	17.6	443

Exhibit C12. Model Estimates for CT Perspectives Scores, Cohort 1 Year 3 Analysis

	Estimate	SE	P-value
Overall impact			
Perspectives	-5.8	3.7	0.117
Belonging	-5.1	4.9	0.296
Interest in Programming	-7.3	5.7	0.202
Engagement	-9.3	4.8	0.052
Motivation to help the world	-4.1	3.6	0.254
Creativity	-4.1	3.8	0.277
Self-efficacy**	-9.2	3.5	0.008
Utility motivation	-2.9	3.9	0.452

*** $p < 0.01$.*

APPENDIX D: VALIDATION ANALYSIS

Test validity refers to the degree to which evidence and theory support the interpretations of test scores for the proposed uses. Contemporary validation analysis is considered as an ongoing process that is initiated at the beginning of assessment design and continues throughout development and implementation. This is particularly important in the case of the CT instruments that support the impact portion of the evaluation. The types of uses and interpretations of the testing results we want to and can make may be different at the different phases of the evaluation (i.e., baseline, midline, and endline), as students experience a range of potential impact of the *CoolThink@JC* lessons at these various timepoints.

In the baseline and midline reports, we provided an overview of the preliminary types of validity evidence that support baseline interpretations of the CT Concepts Level 1 and level 2, CT Practices, and CT Perspectives scores, including aspects of test content and internal structure. We described how well the tests represent the domain of interest using the Evidence-Centered Design (ECD) approach to support test content validity of all three CT instruments. ECD is an assessment design approach that focuses the developer on three questions: What do we want to measure? What evidence is needed? How can tasks be structured to obtain that evidence? The use of this type of process helps ensure that there is alignment between the goals of the assessment and the tasks included in the assessment (Mislevy, 2007).[2] In the earlier reports, we discussed the internal structure aspect of validity for CT Concepts and CT Perspectives by examining test reliabilities and factor structures.

In this appendix, we present validity evidence collected from the endline administration and compare the results to evidence from the baseline and midline administration in order to build coherent validity arguments. The types of validity evidence presented include aspects of test content, internal structure, item validity, and external criteria.

CT Concepts

The internal structure of CT Concepts Level 1, Level 2, and Level 3 forms were examined for the baseline, midline, and endline administrations. We conducted reliability analysis and confirmatory factor analysis (CFA) to see if the items in a test form measure the one CT Concept construct stably and consistently. While there are 4-5 concepts being measured in the CT Concepts test, the number of items for each concept was small and our belief is that these concepts are related enough that we can consider this test as measuring just one construct (students' knowledge of CT concepts). Exhibit D1 presents coefficient alpha for the reliability analysis and fit indices for a one-factor CFA model for CT Concepts. The coefficient alpha for CT Concepts forms ranges from .40 to .66 across time and levels of the lesson sequence. The relatively low reliabilities at baseline were expected because students had not yet started, or had just started, learning the content. Since students were not expected to have a good grasp of the concepts, we expected that they would be doing a large amount of guessing, which

[2] Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463–469. https://doi.org/10.3102/0013189X07311660

can reduce the measured reliability. In addition, the test is intentionally designed to include a range of item difficulties, with some items containing content students would not have seen until a higher level of the lesson sequence. While this range can help us differentiate students, it also can reduce the reliability, as the way students interact with the items is not expected to be similar across all items. Therefore we deemed that the reliability is acceptable for this situation. The floor effect, due to a lack of student knowledge prior to *CoolThink@JC*, also led to the lack of meaningful covariance among items for CT Concepts at baseline. As learning progressed, coefficient alpha increased significantly at midline and endline, despite the smaller numbers of items in tests.

Results from factor analysis show that the hypothesis of items measuring one CT Concepts construct is better supported by the midline and endline data than those at baseline, possibly due to the same floor effect. For Level 1, Level 2, and Level 3 at midline and endline, the one-factor model shows acceptable fit to data as indicated by the fit indices. Examination of the common items across time shows better student performance and greater covariances among these items from baseline to midline to endline, which we would expect as more students are able to answer some of the questions correctly.

Exhibit D1. Summary of Reliability Factor analysis for CT Concepts Forms

CT Concepts	# of	# of	Coefficient	Fit for A One-Factor Model**			
Form	Students	Items	Alpha*	Chi-Square	df	RMSEA	CFI
2017 Feb Level 1	289	31	.40	601.296***	434	.037	.519
2017 Jun/Sep Level 1	439	31	.41	684.076***	434	.036	.649
2018 Jun Level 1	16,594	16	.56	7069.713***	104	.064	.890
2018 Jun Level 2	2,355	21	.63	1627.836***	189	.057	.816
2019 Jun Level 1	5,698	16	.50	2276.200***	104	.061	.884
2019 Jun Level 2	9,412	21	.66	6939.003***	189	.062	.827
2019 Jun Level 3	1,837	20	.48	1323.646***	170	.061	.809

* One common accepted cutoff for good reliability is .70. However, as described in the text above, this cutoff is greatly affected by test length, content domain, the number of constructs measured, and the intended use of a test.

** Common cutoffs for CFA model fit indices: equal to or greater than .90 and .95 for CFI is deemed as good and excellent fit; equal to or smaller than 0.01, 0.05, and 0.08 for RMSEA is deemed as excellent, good, and mediocre fit.

****p* < .001*

We examined item validity by studying item difficulty, discrimination, and missing rate. In each form of CT Concepts, items exhibit a distribution of difficulty levels as intended. Items with more difficult features (for example, those that represented Level 2 content rather than Level 1, or items that contained more intricate code) were found to have a lower item correctness rate. The average correctness rate for CT Concept items increased from baseline (33% and 32% for Level 1) to midline (50% and 44% for Level 1 and Level 2 respectively) to endline (56%, 50%, and 56% for Level 1, Level 2, and Level 3 respectively), indicating the tests were set at appropriate difficulty levels. Exhibit D2 presents the correctness rates for the common items across the different test forms for CT Concepts. With few exceptions, students had increased or stable performance from baseline to midline and endline. Although student performance did not increase from baseline to midline on item q121, it increased significantly at endline. Student performance increased on item q221 from baseline to midline, but dropped a little at endline. This may be due to a reduced focus on conditionals in the Level 2 lessons.

Exhibit D2. Percent Correct for Common Items across CT Concepts Forms, Part 1

Sub-Construct	Level	Item	2017 Feb Level 1 (*N*=289, *J*=31)	2017 Jun/Sep Level 1 (*N*=439, *J*=31)	2018 Jun Level 1 (*N*=16,594, *J* =16)	2018 Jun Level 2 (*N*=2,355, *J*=21)
Repetition	1	q111	37.50%	37.27%	48.97%	51.31%
	2	q121	17.58%	14.42%	-	15.78%
	2	q122	46.64%	44.55%	48.64%	-
Conditionals	1	q212	29.33%	28.17%	35.54%	39.91%
	1	q213	15.11%	14.82%	22.24%	-
	2	q221	26.24%	27.31%	35.32%	31.59%
Parallelism and Sequencing	1	q312a	37.94%	31.24%	40.06%	-
	1	q312b	59.71%	53.38%	54.55%	-
	2	q321	36.07%	38.55%	37.01%	38.31%
	2	q322	33.94%	31.85%	37.76%	38.02%
Data Structure and Algorithm	1	q411a	45.71%	49.88%	74.47%	62.45%
	1	q411b	44.93%	44.37%	67.68%	56.63%
	1	q411c	35.53%	36.38%	45.78%	43.30%
	3	q423a	-	-	80.60%	77.90%
	3	q423b	-	-	78.08%	76.46%
	3	q423c	-	-	65.08%	68.19%

Exhibit D2. Percent Correct for Common Items across CT Concepts Forms, Part 2

Sub-Construct	Level	Item	2019 Jun Level 1 (*N*=5,698, *J*=16)	2019 Jun Level 2 (*N*=9,412, *J*=21)	2019 Jun Level 3 (*N*=1,837, *J*=20)
Repetition	1	q111	56.71%	61.94%	67.21%
	2	q121	-	23.55%	-
	2	q122	-	-	-
Conditionals	1	q212	44.67%	48.46%	-
	1	q213	29.88%	-	-
	2	q221	30.38%	25.21%	26.50%
Parallelism and Sequencing	1	q312a	46.25%	-	-
	1	q312b	63.64%	-	-
	2	q321	37.94%	40.56%	40.16%
	2	q322	37.87%	41.51%	-
Data Structureand Algorithm	1	q411a	84.68%	77.51%	76.10%
	1	q411b	78.11%	73.17%	70.87%
	1	q411c	49.92%	51.91%	52.22%
	3	q423a	83.87%	84.99%	87.19%
	3	q423b	86.53%	87.66%	90.23%
	3	q423c	73.37%	79.19%	82.52%

Across the forms, the overall test discrimination increased from 0.33 and 0.34 (level 1) at baseline to 0.62 and 0.64 (level 1 and level 2 respectively) at midline and 0.57, 0.71, and 0.50 (levels 1, 2, and 3 respectively) at endline, as estimated by a 1PL IRT model. The results indicate that the tests were gaining greater power in differentiating students with different ability levels as they advanced in the pilot lessons.

CT Practices

The design and development of CT Practices follows the ECD approach. Additionally, a set of items was reviewed by having eight students talk out loud as they worked on the items. The students' comments were recorded and reviewed to determine if students were approaching the items as expected and to identify any difficulties interpreting the items. Revisions were made to clarify the items based on the feedback.

Based on reliability analysis for CT Practices, the estimated coefficient alpha is 0.61 at midline and 0.78 at endline. A one-factor model and a four-factor model were fit to data to examine the internal structure of the test. Exhibit D3 shows the factor analysis results for these two models. The one-factor model fit adequately to the data, according to commonly adopted fit criteria.

Exhibit D3. Summary of factor analysis for CT Practices Validation Form for Endline Administration (N=1,808, J=26)

Factor Model	Chi-Square	df	RMSEA	CFI
One-factor model	1386.751***	299	.045	.900
Four-factor model	1079.168***	293	.039	.928

***$p < .001$

Fit of the four-factor model is comparable to the one-factor model. Correlations between the sub-constructs are high, ranging from .62 to .84 at midline and from .68 to .86 at endline. Such results indicate that the use of a one-factor model is appropriate for CT Practices and that the use of one score for the test is appropriate.

Two alternative forms were purposely designed for the validation of CT Practices because of the large number of items included in this test. Students were randomly assigned to one of the two forms—form F and form G—which had the exact same set of items but differ in the item order. Midline analysis found that item correctness rate and item missing rate were related to the item's position in the test: the same item tended to have a lower correctness rate and a higher missing rate if it showed up later in the test. This order effect led to Form F having better overall performance, a better reliability estimate, and higher discrimination power than Form G. This finding is consistent with the earlier finding that students were speeding through the test at midline, which would necessarily result in poor measures of correctness.

For endline administration, modifications to procedures were implemented in order to improve student motivation as they took this test. For example, students were provided with extra paper to work out the problems on, and teachers were asked explicitly to encourage students to do their best. Exhibit D4 presents the percent of max possible scores for CT Practices items in different validation forms at endline. Order effect still exists to some extent, but is not as visible as in the midline results.

CT Perspectives

Earlier reports discussed the evaluation of the seven sub-constructs according to several alternative conceptual models that describe possible relationships among CT Perspectives. The conclusion of these analyses was that most of the items reliably measure their intended subconstructs, and none of the conceptual frameworks fit significantly better than the others (see Shear et al., 2019). Across time, all sub-constructs are highly correlated with each other, except for the construct "Belonging."

Exhibit D5 presents the correlations between the sub-constructs at midline as an illustration. As discussed in the analytic methods section of the main report, we generated a composite CT Perspectives measure by combining the six subconstructs that are highly correlated with each other, namely digital self-efficacy, utility motivation, motivation to help the world, creativity, engagement. The reliability coefficient for the composite measure is very high at 0.95.

Exhibit D4. Percentage of Max Possible Scores for CT Practices Validation Forms for Endline Administration

Sub-construct	Item		Max Possible Score	All Forms (N=1,808, J=26)	Form F (N=935, J=26)		Form G (N=873, J=26)	
					Order	Max %	Order	Max %
Algorithmic Thinking	q11a	Robot moving with blocks	1	56.64%	1	63.94%	23	48.57%
	q11b		1	50.68%	2	60.63%	24	39.79%
	q11c		1	51.10%	3	61.00%	25	40.31%
	q11d		1	40.30%	4	44.43%	26	35.79%
	q12aa	Analyzing the maze	1	42.56%	16	43.30%	7	41.77%
	q12ab		1	32.29%	17	34.29%	8	30.18%
	q12ba		1	26.19%	18	26.76%	9	25.58%
	q12bb		1	26.90%	19	25.03%	10	28.89%
	q13	Analyze the maze part 2	1	42.75%	20	44.10%	11	41.32%
Reusing and remixing	q215	Remaking a picture	1	46.65%	22	21.31%	1	24.03%
	q217		1	48.08%	23	44.80%	2	48.57%
	q21_ mod		2	22.64%	24	42.34%	3	54.03%
	q22a	Making shapes	1	25.31%	6	23.54%	19	27.26%
	q22b		1	19.88%	7	23.37%	20	16.05%
	q22c			16.40%	8	15.63%	21	17.23%
Testing and Debugging	q32a	Painting a picture	1	29.23%	9	34.61%	17	23.32%
	q32b		4	30.91%	10	35.76%	18	25.73%
	q34	Calculator testing	1	29.11%	15	28.07%	12	30.21%
Abstraction and Modularizing	q42	Picking a picture	1	22.06%	5	24.33%	22	19.61%
	q43a	Wizard/ Mandy story	2	65.84%	13	67.90%	13	63.62%
	q43b		2	32.85%	14	33.02%	14	32.67%
	q44a	Average of scores	1	25.13%	25	23.80%	4	26.52%
	q44b			15.69%	26	11.96%	5	19.61%
Testing and Debugging	q51	Robot making cakes	2	46.33%	21	52.01%	6	40.33%
Abstraction and Modularizing	q52a	Playground design	2	70.29%	11	68.75%	15	71.93%
Algorithmic Thinking	q52b	Playground design	5	82.44%	12	83.35%	16	81.45%

Exhibit D5. Correlations between Sub-constructs of CT Perspectives for Endline Administration (*N*=16,341)

	Belonging	Interest	Engagement	Motivation World	Creativity	Self-Efficacy
Interest	.49***					
Engagement	.49***	.93***				
Motivation World	.45***	.79***	.82***			
Creativity	.49***	.83***	.87***	.87***		
Self-Efficacy	.44***	.81***	.82***	.76***	.80***	
Utility Motivation	.46***	.83***	.85***	.84***	.85***	.85***

****p* < .001*

SRI Education™

SRI Education, a division of SRI International, is tackling the most complex issues in education to identify trends, understand outcomes, and guide policy and practice. We work with federal and state agencies, school districts, foundations, nonprofit organizations, and businesses to provide research-based solutions to challenges posed by rapid social, technological and economic change. SRI International is a nonprofit research institute whose innovations have created new industries, extraordinary marketplace value, and lasting benefits to society.

Silicon Valley
(SRI International headquarters)
333 Ravenswood Avenue
Menlo Park, CA 94025
+1.650.859.2000
education@sri.com

Washington, D.C.
1100 Wilson Boulevard, Suite 2800
Arlington, VA 22209
+1.703.524.2053

www.sri.com/education

Stay Connected