CPR2: Fall 2021 Findings

Prepared by:
SRI International
Andrea Beesley
Jared Boyce
Patrik Lundh
Carol Tate
Mindy Hsiao
Elise Levin-Guracar

SRI Education™
A DIVISION OF SRI INTERNATIONAL

SRI International • 333 Ravenswood Avenue • Menlo Park, CA 94025-3493 • 650.859.2000 • www.sri.com

This material is based upon work supported by a grant from the National Science Foundation (#1933678). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Contents

CPR2: Fall 2021 Findings

This memo includes results of data collection from the second cohort of **Collaborative Partnership to Teach Mathematical Reasoning Through Computer Programming (**CPR2) teachers from classroom implementation of CPR2 in fall 2021 and builds on the spring 2021 and Summer Institute 2021 memo.

Overall, fall 2021 teachers were able to implement the CPR2 activities, and students had opportunities to program. However, little support was given to students for exploring mathematical concepts in depth or generating and discussing mathematical conjectures. Teachers adhered closely to the materials they received and most discussion was dominated by teachers' questions and reasoning.

On the student pre-assessment, the treatment and control samples did not achieve baseline equivalence on the pre-test based on What Works Clearinghouse (WWC) standards, meaning students' pre-test survey scores across the two conditions are not similar enough to support a meaningful comparison of CPR2's effectiveness at improving students' mathematical generalization skills.

The Student Computer Science Attitude Survey functioned well; as expected, four of the five factors displayed good reliability (Confidence, Interest, Usefulness, Encouragement). The treatment and control samples achieved baseline equivalence on four factors (Confidence, Interest, Usefulness, Encouragement) based on WWC standards, meaning students' pre-test survey scores across the two conditions were similar enough to support a meaningful comparison of CPR2's effectiveness at increasing students' attitudes about computer programming for these four factors.

We again used the Learning Mathematics for Teaching (LMT) as the teacher pre-assessment. The treatment and control samples achieved baseline equivalence based on WWC standards, meaning teachers' LMT pre-test scores across the two conditions were similar enough to support a meaningful comparison of CPR2's effectiveness at increasing teachers' mathematical generalization knowledge.

Recommendations

Opportunities for the spring semester of CPR2 included:

1. Supporting teachers in discussing mathematical generalization and the connection between the programming activities and generalization.
2. Encouraging teachers to promote student questions and student-centered discussions during CPR2 lessons, and to provide descriptive feedback on students' work.

Study Sample Tracking

This section provides the current state of the teacher sample. In spring 2021, the University of North Alabama (UNA) recruited 49 teachers. After randomization, there were 25 treatment and 24 control teachers. At the time of this writing (December 15, 2021), 24 teachers remained in the study (14 treatment, 10 control; see Table 1). Teachers were attritted for a variety of reasons: they could not attend the Summer Institute (treatment teachers), they did not complete teacher or student baseline data collection, they were reassigned to an ineligible grade, or they took another position in their district that did not involve classroom teaching.

Table 1. Teacher sample after randomization and as of December 15, 2021

	Treatment	Control
Recruited	25	24
Retained	14	10

Lesson Observations

Background: The purpose of the fall lesson observations was to describe how teachers who attended the 2021 Summer Institute implemented CPR2. We analyzed the ways they implemented the CPR2 lessons and supported student engagement and learning and determined whether or not students participated in CPR2 lessons in ways that supported CPR2 learning objectives.

Design: We used the same observation protocol as in fall 2020 and spring 2021, which consisted of two parts: 1) time-stamped running notes to document activities, teacher and student talk, and notes about the learning environment and issues relevant to understanding the lesson; and 2) a debrief organized by descriptive categories aligned with the project's constructs table. The debrief categories were based on the CPR2 instructional model and on other aspects of instruction that we believe support the CPR2 instructional model, including facilitating rich classroom discussions that allow for student questions and reasoning, checking for student understanding, and addressing student misconceptions. Observers took running notes on individual lessons and then wrote summaries for each of the debrief categories.

Data Collection & Analysis: Due to COVID-19 restrictions, we conducted lesson observations virtually. Observers were able to see either the front of the classroom (usually the screen and the teacher, although at times we were unable to see the board/screen) or the teacher's desktop. Importantly, observers did not see students and often the audio quality substantially limited the student talk observers were able to hear. Our observation findings therefore do not fully capture aspects of student engagement, teacher-student interactions outside of front-of-class teacher-led activities, or peer interactions.

We observed ten teachers implementing between one and five lessons for a total of 34 observed lessons. Most of the observed teachers taught the Intro to Python and What is Even? lessons. Some of the observed teachers did not teach the What is Odd? and What is Zero? lessons.

To analyze the data, we created a summary debrief for all observations for each teacher. One SRI Education researcher reviewed all debrief categories across all teacher summaries and described themes and/or variations for each debrief category (e.g., what kinds of questions did teachers ask students, or to what degree did teachers provide student opportunities to write general expressions to represent the mathematical relationships they discovered?). The findings below summarize the themes and variations we saw across all teacher observations.

Findings

Overall, with some exceptions, teachers provided students opportunities to program. Teachers generally seemed to feel confident in the CPR2 content. In the lessons we saw, there was little support for exploring mathematical concepts in depth or generating and discussing mathematical conjectures. Teachers tended to follow the "script," with an emphasis on students correctly following directions. Teachers' questions and reasoning dominated instruction and whole-class discussions. There was little support for checking and adapting to student understanding.

Presence of CPR2 Instructional Model: Most teachers provided opportunities for students to write mini-programs and to write general expressions. However, there appeared to be an emphasis on "doing" and following instructions, with little opportunity for students to explore mathematical concepts through the programming and virtually no chance for students to write and explore mathematical conjectures based on what they were learning through the programming. Teachers also generally did not identify and explain the key concept of generalization. One teacher provided more student opportunity by having students work on debugging programs as well as writing their own programs (to add another column of even numbers) from scratch. Four teachers provided time for students to program, but the activity was scripted, with students copying code rather than exploring on their own. One teacher gave students time to write general expressions to represent even and odd numbers, but also made time to discuss other ways to write expressions for even and odd numbers.

Teacher Capacity: Most of the teachers appeared to be confident in, and have a good grasp of, the CPR2 content. For example, one teacher knew the general expressions and what to ask when students were not sure about general expressions for even or odd; this teacher also guided them to 2n and 2n + - 1 and knew how to debug student work. Another teacher seemed confident with the topic and taught it in a way that seemed to be paced well for students. The lessons were not rushed, and students were engaged and participating (seemingly equally) because it appeared they were all on the same page before moving on. Other teachers occasionally struggled. For example, one teacher seemed to know the programming but did not

know the compiler very well. At one point, they asked the observer to help. Eventually, a student figured it out. Another teacher appeared confident when talking about content, while they frequently struggled with the debugging, such as indenting in the loop, or realizing that n<5 would not print a column of 1-5 but rather 1-4.

Instructional Practices: Most teachers did not clearly frame the lessons or connect activities to prior learning. In their instruction, teachers tended to "follow the script" with little student input into the content of activities and discussions. For several teachers, compliance and following teacher directions were either explicitly or implicitly expected. Some teachers encouraged students to speak up and to ask questions. A couple of teachers were exceptions. One of them, who generally seemed more flexible and responsive to students, gave students ample time to explore their programs and come to an understanding of the program. They also asked questions and gave students time to think and answer. But overall, teachers did most or all of the "cognitive work" in the classrooms observed. This meant the teacher was primarily the one to ask questions, initiate discussions, verbalize their reasoning, and do the explaining, with varying degrees of student input. For example, one teacher drove the reasoning with some student input; this teacher did most of the problem solving, explaining, and reasoning. When debugging, the teacher did very little to clarify what they were doing and why. Instead, they simply debugged the program and moved onto the next student. Given the virtual observation format, it was difficult for observers to gauge student engagement. We did note that students appeared engaged or interested in about half the classes, and more passive and mostly following directions in the other half.

Whole-class Discussion and Teacher Questions: The common instructional approach of teachers owning the reasoning and students following directions was reflected during whole-class discussion as well, which can otherwise be an opportunity to surface student questions, struggles, and ideas. During whole-class discussions, about half the teachers tried to respond to and build on students' ideas. The discussions were primarily led and informed by the teachers' questions. One teacher built on student answers in the style of "What else are we missing?" The other half of classes were dominated by teacher reasoning. For example, one teacher drove the conversation toward answers they were already looking for. Students answered correctly, and the teacher explained the reasoning rather than having the students explain. In all but two classrooms, teachers primarily asked "fill-in-the-blank" and funneling questions, rarely using open-ended questions to prompt students' own reasoning. In a couple of cases, observers noted that teachers would quickly answer their own questions without giving students enough time to think and respond. Two teachers, while also relying primarily on fill-in-the-blank questions, did include more open-ended ones as well. One of them, while the class was exploring arithmetic operators, had students explain what they noticed when they printed each one. The teacher did the same for equal to/less than/etc. This teacher also had students explain why they thought each operator did what it did in code.

Teacher Support for Student Understanding: Teachers' practices for checking student understanding primarily took the form of call-and-response or walking the room and checking student work. Across all classrooms (except one in which student questions were not audible), students either asked no questions or only asked questions about procedure. No students were observed asking conceptual questions. Teachers provided little descriptive feedback to students. There were some instances of praise and encouragement. One teacher, while not providing feedback *per se*, did continue to build on students' responses, asking follow-up questions to explore their reasoning. Given that teachers did not do much checking for understanding, there were few adaptations to instruction in response to student thinking. Most teachers appeared focused on going through the slides with little or no change.

Student Assessment Analysis

Background: One of the goals of CPR2 is to "increase student performance in problems involving [mathematical] generalization." We intended to measure students' mathematical generalization skill during the efficacy study using a student assessment specifically designed to measure this skill.

Design: In the 2021 iteration of the student assessment, we retained two of the Mathematics Assessment Resources Services (MARS) items from the 2020 assessment and added seven released items from the National Assessment of Educational Progress (NAEP) 4th and 8th grade math assessment. We searched for items under "number properties and operations" and "algebra" for the content areas and, where available, "conceptual understanding" and "problem solving" for ability. The UNA and SRI teams reviewed the items to ensure the questions were a reasonable match with the CPR2 program, and of 11 reviewed items, nine were retained for the student assessment. The NAEP provides performance data for their released questions,[1] and we used these to determine if the questions were at an appropriate difficulty level for the sample. Additionally, we piloted these items in spring 2021 and found they performed well for our purposes.

Data Collection: Students in both treatment and control classes took the student assessment as a part of our single data collection instrument. Students completed the assessment prior to any teacher implementation of CPR2 (in treatment classes). In control classes, teachers had students complete the assessment by September 30, 2021. Students completed the assessment prior to taking the pre-survey. We obtained responses from 574 treatment students and 472 control students who completed the survey and consented to have their data used for research purposes.

Findings

We conducted two analyses on the student pre-test data: descriptive statistics (overall and by item) and baseline equivalence testing in the style of WWC. The overall scores are reported in Table 2 below and the by-item analysis is in Table 3.

The mean treatment score was 2.68 out of 9 points with a standard deviation of 1.73, while the mean control score was 3.6 with a standard deviation of 2.26 (see Table 2). There is substantive room for student learning growth over the academic year, providing sufficient response space to measure a potential impact of CPR2 on treatment students' mathematical generalization performance as well as the business-as-usual growth of control students from their mathematics coursework.

The treatment and control samples did not achieve baseline equivalence on the student assessment pre-test based on WWC standards, meaning students' pre-test survey scores across the two conditions were not similar enough to support a meaningful comparison of CPR2's effectiveness in improving students' mathematical generalization skills. Baseline equivalence tests are reported as effect sizes, and the effect size for this pre-test was -0.46, which is outside of the +/- 0.25 guidelines set by WWC. Control students outperformed treatment students on their overall scores and on seven specific test items.

Since the two groups were not equivalent at baseline, any potential impacts of CPR2 on student learning will need to be interpreted conservatively. For example, any positive impacts of CPR2 may not be due to the strength of the CPR2 program but rather to treatment students having more room for growth on the assessment. Similarly, any negative impacts of CPR2 may not be due to a flaw in the CPR2 program but rather to control schools having stronger math curriculum and instruction.

For this midyear analysis, we analyzed the student assessment pre-test data without accounting for students being clustered within teachers with a teacher-level treatment design. The impact analyses at the end of the year will use hierarchical linear modeling (HLM) to account for clustered data.

Table 2. Student pre-test results, treatment, and control

		Pre-test (out of 9 points)
Treatment	Mean	2.68
	SD	1.73
	n	551
Control	Mean	3.60
	SD	2.26
	n	460
Baseline Equivalence		-0.46

Table 3. Student pre-test results by item, treatment, and control

Item Analysis	Treatment Percent correct	Control percent correct	*p*-value
Item 1	22%	30%	<0.01
Item 2	32%	35%	0.30
Item 3	32%	38%	0.06
Item 4	26%	38%	<0.01
Item 5	15%	25%	<0.01
Item 6	22%	38%	<0.01
Item 7	41%	53%	<0.01
Item 8	46%	58%	<0.01
Item 9	32%	44%	<0.01

Note: Treatment n = 551, Control n = 460, see Appendix A for full test questions and response options.

Student Survey Analysis

Background: One of the research goals of the CPR2 study is to increase "the extent to which students feel comfortable with the programming activities and with the associated mathematics, [and] the extent to which they would be interested in similar activities in the future." We intended to measure students' comfort and interest in programming activities during the efficacy study through a student survey.

Design: We based our student survey on the Student Computer Science Attitude Survey[1]. This survey was tested and validated in 2010–2016 for grade 8+ students for measuring five attitudinal constructs related to computer science:

- Students' **confidence** in their ability to learn computer science skills and solve computer science problems.
- Students' **interest** in learning computer science and solving problems.
- Students' perceptions of **belonging** in computer science.
- Students' beliefs in the **usefulness** of learning computer science.
- Students' perceptions of being **encouraged** to study computer science.

In fall 2020, in partnership with the UNA and Horizon Research, Inc. (HRI) teams, we reviewed the survey items and concluded they were appropriate for measuring the study objectives. During the 2020-21 school year, we piloted the student survey. We found the instrument to have acceptable internal consistency reliability, and that students responded meaningfully to the items. Four of the survey factors had good reliability with our student population: Confidence, Interest, Usefulness, Encouragement. Our piloting data indicated that one factor, Belongingness, did not perform as well with middle school students. We decided to keep the

[1]Haynie, K.C. and Packman, S. (2017). *AP CS Principles Phase II: Broadening Participation in Computer Science Final Evaluation Report*. Prepared for The College Board and the National Science Foundation, February 12, 2017. Skillman, NJ.

survey instrument intact (i.e., in the same form used in prior validation research by its creator) rather than remove the Belongingness survey items. See Appendix B for full reliability testing information.

Data Collection: Students in both treatment and control classes took the student survey as a part of our single data collection instrument for students. Students completed the survey after taking the pre-assessment. We obtained responses from 445 treatment students and 398 control students who completed the survey and consented to have their data used for research purposes.

Findings

We conducted three analyses of the student survey pre-test data: reliability testing, descriptive statistics, and baseline equivalence tests in the style of WWC. The results of the descriptive and baseline equivalence analyses are reported in Table 4 below.

Our reliability testing of the student survey pre-test data confirmed our findings from the 2020–21 survey pilot. As before, four of the factors performed well (Confidence, Interest, Usefulness, Encouragement) and one did not (Belongingness). A more complete description of reliability testing can be found in Appendix B.

Each item was measured on a 1-4 Likert-style scale (1 = strongly disagree, 4 = strongly agree) for an overall factor score range of 5-20 (12.5 midpoint). Overall, students reported either modest disagreement or mixed agreement across all five factors. All factors averaged between 10 and 13, which is an average item score between 2.0 (disagree) and 2.6 (in between disagree and agree).

The treatment and control samples achieved baseline equivalence on four factors (Confidence, Interest, Usefulness, Encouragement) based on WWC standards, meaning students' pre-test survey scores across the two conditions were similar enough to support a meaningful comparison of CPR2's effectiveness at increasing students' attitudes about computer programming for these four factors. Baseline equivalence tests were reported as effect sizes, with these four factors having baseline differences between 0.03 and 0.20 across the two conditions. These four values are within the +/- 0.25 guidelines set by WWC.

One factor, Belongingness, did not achieve baseline equivalence. Belongingness had a baseline difference of 0.26, which is above the +/- 0.25 guidelines set by WWC. We are unconcerned with this significant baseline difference for two reasons. First, this factor barely achieved acceptable reliability in our pre-test data ($\alpha = 0.703$) and did not perform well in our pilot test ($\alpha = 0.650$). Second, based on the poor reliability from our pilot, we contacted the researcher who created and validated the Student Computer Science Attitude Survey to discuss our findings. She was unsurprised that Belongingness would be less reliable for our population (middle school students), and she was not certain what this factor would represent in this age range or for an intervention like CPR2. Accordingly, we decided not to interpret this factor in our findings prior to conducting the baseline equivalence test, and the lack of baseline equivalence provides additional evidence in support of this decision.

In this midyear analysis, we analyzed the student survey pre-test data without accounting for students being clustered within teachers with a teacher-level treatment design. The impact analyses at the end of the year will use hierarchical linear modeling (HLM) to account for clustered data.

Table 4. Student pre-survey results by construct, treatment, and control

		Confidence	Interest	Belongingness	Usefulness	Encouragement
Treatment	Mean	12.1	11.6	12.3	12.9	10.6
	SD	2.88	3.41	2.70	3.03	3.12
	N	445	442	445	441	443
Control	Mean	11.5	11.3	11.6	12.3	10.2
	SD	3.02	3.34	2.63	3.16	3.06
	N	398	398	398	398	398
Baseline Difference		0.20	0.09	0.26	0.19	0.13

Note: Factors were only assessed for a student if they replied to all five items for a factor, which is why the N's may vary across factors.

Learning Mathematics for Teaching (LMT)

Background: The LMT assessment was developed by the University of Michigan to measure teachers' understanding of the mathematical pedagogical and content knowledge teachers need to teach mathematics well. We selected items from the Middle School Patterns, Functions, and Algebra content area of the LMT to measure whether CPR2 improves middle school mathematics teachers' understanding of generalizability and patterns.

Design & Analysis: During the 2019-20 pilot year, we developed and piloted a pre- and post-test using items from the LMT Middle School Patterns, Functions, and Algebra content area and balanced the two tests on difficulty, total number of items, content, and types of questions. We focused our item selection on questions related to mathematical generalization that were well-aligned to CPR2 content in general without being overly aligned to the specific CPR2 learning activities. Our piloting process indicated the items we selected and the form designs overall worked well for our purposes. Thus, we used our piloted forms for the current year. We consulted with both UNA and Horizon Research in item selection, item balancing across tests, and interpretation of the piloting results.

We ran a two-parameter item response theory (IRT) analysis using the difficulty and discrimination values for each item that were provided by the test developers based on prior data. The IRT analysis estimates both item parameters and ability estimates. We used the parameters item difficulty and item discrimination for the IRT model. The item parameters set the scale and are used to estimate ability. This type of model accounts for how challenging items are, versus more traditional models that give equal weight to each item. The analysis returns

estimates of teachers' results expressed as standard scores that can be readily compared to one another.

Data Collection: We administered the LMT to both treatment and control teachers as both a pre-test before the first day of professional development (PD) (for treatment teachers). In total, 22 treatment teachers and 17 control teachers completed the LMT. Of those teachers, 14 treatment teachers and 10 control teachers have remained in the study.

Findings

We conducted three analyses of the LMT pre-test data: descriptive statistics, baseline equivalence tests in the style of WWC, and attrition *t*-tests. The results of these analyses are reported in Table 5 below.

We calculated the mean standardized teacher LMT score both for the original set of teachers (all of the teachers who took the LMT prior to attrition), and for the current sample (the teachers who remained in the study as of December 15, 2021. The mean standardized teacher LMT score for the original sample (n = 22) was .15 and the control mean for the original sample (n = 17) was .04. With the small sample sizes for each condition, the standard deviations (.82 for the original treatment and .75 for the original control groups) were quite high relative to the mean difference. The means of the current samples were closer together (0.20 for current treatment sample, 0.21 for current control sample).

The treatment and control samples achieved baseline equivalence based on WWC standards, meaning teachers' LMT pre-test scores across the two conditions were similar enough to support a meaningful comparison of CPR2's effectiveness at increasing teachers' mathematical generalization knowledge. Baseline equivalence tests are reported as effect sizes, with the original samples having a baseline difference of 0.14 and the current samples having a baseline difference of -0.01. Both of these values were within the +/- 0.25 guidelines set by WWC.

The attrition *t*-tests did not identify any statistically significant differences between the original and current treatment samples (p = 0.84) or the original and current control samples (p = 0.59). This suggests teacher attrition did not significantly change the mean values of the LMT pre-test for either condition. However, this could be more reflective of the small sample sizes than a lack of change in the mean scores due to attrition, with the *t*-tests possibly being underpowered to detect such a change.

Table 5. LMT Pre-test results, treatment and control, original and current samples

		Original sample	Current sample	Attrition *p*-value
Treatment	Mean	0.15	0.20	0.84
	SD	0.82	0.84	
	n	22	14	
Control	Mean	0.04	0.21	0.59

	Original sample	Current sample	Attrition *p*-value
SD	0.75	0.79	
n	17	10	
Baseline Difference		0.14	-0.01

Spring 2022 Upcoming Activities

In spring 2022, we will be wrapping up data collection by conducting virtual classroom observations for the treatment teachers followed by a post-lesson implementation questionnaire, student post-survey, student post-assessment, teacher post-assessment with the LMT, and teacher interviews.

Spring 2022 Teacher Data Collection

Treatment and control teachers will complete the post-assessment of the LMT before the end of the school year, no earlier than April 18, 2022, and only after all CPR2 instruction has concluded (for treatment teachers).

We will continue to administer teaching logs across both conditions. For treatment teachers, we will ask them to complete a teaching log shortly after they complete their spring CPR2 instruction to document what content and activities they implemented. For control teachers, we will ask them to complete a semester overview teaching log near the end of spring semester to identify whether control students may have received instruction similar to CPR2.

We will observe treatment teachers on their delivery of CPR2 lessons. We will plan to use the same protocol to observe and analyze the spring even-odd-consecutive lessons. We plan to observe eight to 10 teachers. We will be asking all treatment teachers to complete a post-lesson implementation questionnaire about their experiences teaching the spring lessons.

We plan to interview approximately half of the treatment teachers following their spring implementation. Topics of the interview will include their confidence regarding integrating programming activities into their classes, their perceptions of student learning and engagement, and their thoughts about how programming activities might support the development of mathematical generalization abilities. We will also ask teachers about any plans they have for using CPR2 activities in the future.

Spring 2022 Student Data Collection

Students of both treatment and control teachers will complete post-surveys and post-assessments this spring. Treatment teachers will administer the post-survey and post-assessment to their students after teaching their spring CPR2 lessons and before the end of school, though no earlier than April 18, 2022. Control teachers will administer the post-survey and post-assessment to their students between April 18, 2022 and their last day of school.

Appendix A: Student Assessment Items

Item #1 (multiple choice): "If *n* is any integer, which of the following expressions must be an odd integer?"

Response Options
n+1
2n
2n+1
3n
3n+1

Item #2 (multiple choice): "According to the pattern suggested by the four examples above, how many consecutive odd integers are required to give a sum of 144?"

1 + 3 = 4
1 + 3 + 5 = 9
1 + 3 + 5 + 7 = 16

Response Options
9
12
15
36
72

Item #3 (multiple choice): "If *n* represents an even number greater than 2, what is the next larger even number?"

Response Options
n + 1
2n + 1
2n
n + 2
n2

Item #4 (multiple choice): "Which of the following is always an odd integer?"

Response Options
The product of two odd integers
The product of two consecutive integers
The sum of three even integers
The sum of two odd integers
The sum of three consecutive integers

Item #5 (open response, bounded): "If the product of 6 integers is negative, at most how many of the integers can be negative?"

Response Options
0
1
2
3
4
5
6

Responses for Item #6 (multiple choice): "Which expression is the greatest when *n* is a negative number?"

Response Options
n – 2
2n
n2
n/2
2/n

Item #7 (multiple choice): "A car can seat *c* adults. A van can seat 4 more than twice as many adults as the car can. In terms of *c*, how many adults can the van seat?"

Response Options
c + 8
c + 12
2c – 4
2c + 4
4c + 2

Item #8 (multiple choice): "Each of the 18 students in Mr. Hall's class has *p* pencils. Which expression represents the total number of pencil's that Mr. Hall's class has?"

Response Options
18 + p
18 – p
18 * p
18 / p

Note: Overall *n* = 78.

Item #9 (multiple choice):

□	△
4	9
5	11
6	13
7	15

Which rule describes the pattern shown in the table?

A. □ + 5 = △

B. □ + □ = △

C. □ + □ + 1 = △

D. □ + □ + 2 = △

Correct answer: C

Appendix B: Student Survey Reliability Testing

We ran reliability testing on the current pre-test and the fall 2020 pilot sample to compare to the research done on the Student Computer Science Attitude Survey.[2] This survey was tested and validated across 2010–16 for grade 8+ students for measuring five attitudinal constructs related to computer science. Since the survey was originally validated against mostly high school students, we wanted to check reliability against the study's population of middle schoolers during piloting year before using this survey for the implementation study. We sought to answer three main questions in the pilot:

1. Are the survey factors still reliable with our intended student population?
2. Do students appear to meaningfully respond to the survey?
3. Do we have significant risk of response ceilings or floors such that intervention impacts would be difficult to determine?

The factor structure of the original survey was broadly maintained in the pilot administration. Three of the five factors had very good reliability with our student population, one had good reliability, and one had minimally acceptable reliability.[3] Our reliability metrics for most factors were slightly below those of the original research (see Table below), which may be partially due to the differences in sample sizes. We consider only one factor, Belongingness, to be potentially problematic for our study. Our reliability calculation was barely acceptable (0.650) and significantly below that of the original research (0.850). We addressed our piloting results with the original survey author, and we determined that Belongingness may not be a meaningful construct for our study given the age group of our students and the design of CPR2. We decided not to edit the survey (i.e., in the same form used in prior validation research by its creator) or remove the Belongingness survey items. While removing these items would be unlikely to affect students' responses, we decided to err on the side of caution.

During the fall 2021 administration we again ran reliability testing with this larger middle school sample to confirm the four factors we found reliable in the pilot study performed well with the larger sample. We again found the Confidence, Interest, Encouragement, and Usefulness factors had either very good reliability or good reliability. This further confirmed suitability of the instrument for this group of students. While Belongingness barely achieved good reliability with the larger sample (0.703), we decided not to analyze these data given the lack of reliability in our pilot and the survey author's sense that this may not be meaningful for our age group or for the CPR2 intervention.

---

[2] Haynie, K.C. and Packman, S. (2017). *AP CS Principles Phase II: Broadening Participation in Computer Science Final Evaluation Report.* Prepared for The College Board and the National Science Foundation, February 12, 2017. Skillman, NJ.

[3] We consider Cronbach's alphas of 0.80+ to indicate very good reliability, 0.70+ good reliability, and 0.60+ minimally acceptable reliability per Nunally, J.C. (1967). *Psychometric Theory*. New York: McGraw-Hill.

Reliability Testing

Psychometrics		Confidence	Interest	Belongingness	Usefulness	Encouragement
Current Pre-test	Cronbach's alpha	0.766	0.843	0.703	0.815	0.813
	n	843	840	843	839	841
Pilot Sample	Cronbach's alpha	0.850	0.892	0.650	0.853	0.778
	n	149	149	149	149	149
Research Reference	Cronbach's alpha	0.890	0.932	0.850	0.892	0.858
	n	802	802	803	802	805

Note: Factors were only assessed for a student if they replied to all five items for a factor, which is why the *n*'s may vary across factors.

Appendix C: Student Survey Items by Construct

Construct	Items (1-4 scale; 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree)
Factor 1: Confidence	I am sure I could do advanced work in computer programming.
	I have self-confidence when it comes to computer programming.
	I am confident that I can solve problems by using computing.
	I can learn computer programming without a teacher to explain it.
	I think I will do well in computer programming.
Factor 2: Interest	I like writing computer programs.
	I like to use computer programming to solve problems.
	The challenge of solving problems using computer programming appeals to me.
	I would take additional computer programming courses if I were given the opportunity.
	I hope that my future career will require the use of computer programming.
Factor 3: Belongingness	I feel I belong in computer programming.
	I feel comfortable in computer programming.
	I feel accepted by my peers in computer programming.
	I know a lot of students like me who are interested in computer programming.
	I know someone like me who uses computer programming in their work.
Factor 4: Usefulness	Skills used to understand computer science material can be helpful to me in understanding things in everyday life.
	Computer programming is a worthwhile and necessary subject.
	Knowledge of computer programming will help me earn a living.
	Learning to use computing skills will help me achieve my career goals
	I'll need a mastery of computer programming for my future work.
Factor 5: Encouragement	A friend or peer has encouraged me to study computer programming.
	Someone in my family has encouraged me to study computer programming.
	Someone I know has discussed with me the computer programming field.
	Someone I know has given me the desire to study computer programming.
	Someone I know has given me the desire to study computer programming.
	Someone I know has praised my work in computer programming.