Understanding How Integrated Computational Thinking, Engineering Design, and Mathematics Can Help Students Solve Scientific and Technical Problems in Mathematics and Career Technical Education (INITIATE)

NSF Award #1741784

Evaluation Report

August 2018 – June 2019

Prepared by Acumen Research and Evaluation, LLC

Acumen
Research and Evaluation, LLC
2525 N. Reynolds Road, #3
Toledo, OH 43615
419-260-1811

"Accuracy of observation is the equivalent of accuracy of thinking."---Wallace Stevens

Table of Contents

Executive Summary

Due to the timing of the required project annual reports, the 2018-19 evaluation report focuses on Cohort 1—those who attended the 2018 Summer Institute. Some 2019-20 evaluation data has been collected (student outcomes for Cohort 1, Cohort 2 baseline, and 2019 Summer Institute) and will be analyzed and reported to INITIATE leadership in August and included in the next annual evaluation report.

Six teachers from Bowsher High School in the Toledo Public School district (Ohio) were recruited for cohort 1. Five were Mathematics teachers and one was a Career Technology teacher. Cohort 2 consisted of 17 teachers representing 8 different high schools in the district. Nine were mathematics teachers, five were CTE teachers, and three were special education support teachers working in the math and CTE classrooms. Of the 17 teachers, five were men and five represented minority populations (29% in both cases).

Classroom observations of each teacher teaching what they considered an inquiry-based lesson prior to participation in the INITIATE Summer Institute were conducted in May 2018. Using the Electronic Quality of Inquiry Protocol, observations were rated on four factors of inquiry-based instruction: Instructional, Discourse, Assessment, and Curriculum. These scores were compared with post-Institute observation scores to examine growth and possible influence of the INITIATE project. Teachers showed small to moderate gains across each of the four factors, sometimes moving from a Developing Inquiry to the Proficient Inquiry category overall.

Teacher confidence in adopting the INITIATE Model (a major research question as well as a formative evaluation tool) was measured with the Stages of Concern Questionnaire (May 2018, October 2018, and February 2019). Just prior to participation in the Summer Institute, the teachers ranked in the lowest levels of concern: They were interested in the INITIATE model but were not quite sure it was relevant to their teaching. The October testing showed that teachers moved beyond the informational level and became concerned with management or how implementing INITIATE might fit into the classroom schedule. The February responses revealed little change and continued to reflect a lack of a clear understanding of the INITIATE teaching model. Coupling with classroom observations and teacher focus group interviews verified that the teachers in Cohort 1 did not gain enough understanding of the model to feel comfortable implementing it.

A focus group interview with the teachers on the last day of the Institute gathered qualitative formative/process data that informed positive change for the second Institute (2019). In general, teachers felt the INITIATE team was very organized and well-coordinated; however, they did note that there was some down time that could be eliminated as well as a few sessions that seemed irrelevant to how they will use the self-driving cars in their classroom. A second focus group interview was conducted with Cohort 1 at the end of their participation in the program where teachers provided qualitative feedback regarding their experience with all facets of the INITIATE program. Overall the teachers found the Summer Institute to be a helpful experience. The teachers felt like the goals of the program were not always clearly articulated and this caused some confusion during the academic year implementation (and is reflected in the SoCQ

analysis). Teachers struggled to find ways to incorporate all of the components of INITIATE (inquiry instruction, computational thinking, project-based learning, etc.) into the prescribed class time and felt that the program would be more successful in phases: with teachers learning about and introducing smart cars early, then later on developing an integrated project.

A test of computational thinking was given to the teachers on the first and last days of the Institute as well as eight months after the Institute. A scoring rubric was developed by a team comprised of INITIATE content experts and Acumen measurement experts. It was developed based upon computational thinking best practices. Reliability Initial reliability was established between three members of Acumen and INITIATE. These three raters then scored all of the completed tests. Results showed moderate to large gains in teachers' knowledge of computational thinking after participating in the Summer Institute. These gains were sustained in five of the teachers.

To gather data about student outcomes, students completed three surveys spring 2019—a demographic survey, a student engagement survey, and a student computational thinking survey. The purpose of administration for students in Cohort 1 classrooms was to establish reliability and gather validity evidence. This data is currently being analyzed and results will be shared with INITIATE leadership late summer as well as included in the next evaluation report. Next year the tool will be used to examine student outcomes.

All findings noted above were shared with INITIATE leadership over the course of the last year. INITIATE leadership, upon review, made major revisions to the Summer Institute curriculum including limiting instructional lectures and providing more time for teachers to learn through application. More linkages to mathematics were included. In addition, the expectations for implementation of Summer Institute learning have been clearly defined and provided to teachers. Teachers were aware of these expectations prior to the Summer Institute and they guided learning and lesson development. Teachers did gain in their understanding and application of computational thinking as a result of participation. In summary, INITIATE is evolving towards a replicable model. Year 3 data collection and analysis of Cohort 2 should show improvement.

1) Evaluation Model

The following revised evaluation plan (Table 1) was put in place in 2017-18. No changes were made in 2108-19. Measures written in blue have been administered; those in red have also been analyzed. The remaining measures will be completed during the summer and next year.

Table 1. INITIATE Evaluation Plan

Evaluation Question	Measure	Analysis	Frequency
1	Faculty developed content tests for teacher PD topics	Pre/Post participation comparisons using paired t test. Effect sizes will also be calculated	Prior to and after participation in summer program
2, 3 & 8	EQUIP Stages of Concern Questionnaire (SoCQ)	Pre/Post participation observations with comparisons using Rasch modeling and paired t test; effect size calculations; qualitative analysis of open-ended surveys & interviews, SoCQ scores will be correlated (factor analysis or multiple regression) to explore strength of relationships.	Prior to and after participation in summer program.
4	Meeting observations, teacher assessments of lessons using a project-developed rubric, teacher focus group interviews, EQUIP, SoCQ	Case (observation) notes will be analyzed using qualitative measures; rubrics will be compared with an agreed upon baseline or minimum ranking (analysis with Rasch Measurement Model will provide reliability and validity evidence), EQUIP and SoCQ analyzed using method described above.	Data collected every two months—analyzed quarterly.
5	Ohio mathematics assessments, project-developed student CT assessments, student capstone projects, EQUIP, student advanced science enrollment numbers, data from online classroom platform	Scores will be compared between treatment and control to examine skill attainment (independent t test). Scoring rubrics developed in collaboration with faculty will be used for capstone projects with an established benchmark score; treatment student enrollment numbers in advanced mathematics will be compared with control using chi square. EQUIP will provide supporting evidence (i.e., students will be observed participating in the classroom). Qualitative analysis of data from online platform.	Annually as completed.

6	Project artifacts and records, interviews with staff, observations of Summer PD using Horizon Professional Development Observation Protocol	Implementation will be compared with design. Interviews will delve into discrepancies and the Horizon observation tool will provide a rich description of PD implementation based upon PD best practices.	Quarterly (PD annually).
7	Culmination of answers to evaluation questions 1 – 5		Annually.

The corresponding evaluation questions are as follows:

1. Does INITIATE teacher PD improve teacher understanding of CT standards and PBL?

2. Do participating teachers implement the treatment protocol (i.e., PBL and inquiry-based instruction) with fidelity?

3. Do teachers integrate CT into their mathematics and computer science teaching?

4. Does lesson sharing among mathematics teachers result in lessons that address the intentions of INITIATE? Do teachers use the lessons and are they implemented as designed?

5. Given #1, #2, and #3 are achieved to satisfactory levels, does this indeed improve student CT, student interest in mathematics, and student mathematics ability?

6. Are all project activities offered and implemented as designed? If not, are there justifiable reasons for changes?

7. Based upon findings from answering evaluation questions 1 – 5, does the INITIATE model work?

8. Does the Stages of Concern Questionnaire function as an effective tool for diagnosing and addressing teacher concerns and implementation?

2) Cohort 2 Sample Demographics

The Cohort 1 (2017-18) sample was discussed in the previous annual evaluation report. Based upon feedback from Cohort 1, recruiting for Cohort 2 was expanded to include all secondary school math/career technical education (CTE) teachers within the Toledo Public Schools. As a result, Cohort 2 consisted of 17 teachers representing 8 different high schools in the district. Nine were mathematics teachers, five were CTE teachers, and three were special education support teachers working in the math and CTE classrooms. Of the 17 teachers, five were men and five represented minority populations (29% in both cases). Subjects taught by these teachers included math: algebra, geometry, applied math, and trigonometry; CTE: construction and manufacturing technology, diesel technology, CAD design, medical technology, and precision machining. The content teachers (excluding the special education support) averaged 14 years of teaching experience and all have participated in some form of previous professional development (PD) although in general the PD was school sponsored and short term. Two of the math teachers hold a Master's degree. The Cohort represented a thorough mix of both mathematics and CTE expertise. While some data has been collected on Cohort 2, growth comparisons and findings will be presented in the 2019-20 evaluation report. All remaining sections focus on Cohort 1.

3) Teaching Practice (Cohort 1): Evaluation questions 2, 3, 4, 5, & 8

In May 2018, Acumen visited each of the six participating teachers' classrooms to observe a typical inquiry-based math or computer science lesson. Subsequent observations were made in the academic year following the teachers' completion of the Summer Institute in June 2018. The purpose of the post-Institute observations was to observe teachers delivering one or both of the lessons they developed during the Summer Institute. The evaluation also examined degree to which teachers implemented the INITIATE model as designed. Observations were scored using the Electronic Quality of Inquiry Protocol (EQUIP) (Inquiry In Motion, Clemson University).

The EQUIP rubric measures four factors associated with inquiry instruction and based upon NGSS—instruction factors, discourse factors, assessment factors, and curriculum factors. Within these four factors are 19 indicators. Scores from pre- and post-Institute observations were compared to determine if there are areas of improvement between the observations. The rating scale includes four levels—pre-inquiry, developing, proficient, and exemplary. Researchers have found a strong positive relationship between proficient scores on the EQUIP and student achievement (Marshall & Horton, 2011).

The EQUIP observation categories were converted to a four-point ordinal scale (pre-inquiry became a 1, developing became a 2, etc.) to better analyze the data. Previous evaluations (e.g., NURTURES—University of Toledo NSF MSP) showed the EQUIP observations measure our desired trait (i.e. inquiry facets of instruction).

Of the six teachers, three were observed twice, providing nine total observations. All eight of the observations in the math classes lasted 47 minutes – the normal duration of a class period in the Toledo Public High School district – while the CTE electrical engineering followed a "block" two-period schedule and lasted 94 minutes. It should be noted that many of the lessons were designed as multi-day "units" and the segment evaluated could have been any single part of that unit. Course subjects included algebra (standard and honors), geometry, computer aided design, and college credit level statistics. The classes were comprised of students from across all high school grade levels including three mixed grade classes. On average, the nine classes observed were comprised of 46% female students, 57% minority students, and a class size of 17 (class size ranged from 10 to 24).

There are not enough observations to meet the requirements needed for the reliability of estimates of teacher ability based on item response theory. Therefore the following sections summarize the teachers' performances across the four different inquiry-based factors measured by the EQUIP instrument.

Instructional

Table 2 reports the ratings for teachers' performance on the five constructs measured under the Instructional factor. Prior to participation in INITIATE the teachers were already performing at an acceptable level (Proficient) on this factor. Overall, the median rating showed an increase in the post-Institute observations on the "Order of Instruction" indicator. This indicator assesses the extent to which the teacher allowed students time to explore ideas before providing an explanation and whether the teacher or students provided explanations.

Across all teachers' ratings for Instructional constructs, Pre-Inquiry was selected only twice, compared to the four times this category was selected on the pre-Institute observations. The Exemplary Inquiry rating was selected seven times, compared to three occurrences pre-Institute, with one teacher receiving Exemplary designation on three out of the five Instructional factor indicators.

Table 2. Comparison of Cohort 1 observations scores for Instructional Factors

		Instructional Strategies	Order of Instruction	Teacher Role	Student Role	Knowledge Acquisition
Pre-Institute	Mode	3	1, 3	3	3	3
	Median	3	2.5	3	3	3
Post-Institute	Mode	3	3	3	3	3
	Median	3	3	3	3	3

Discourse

Table 3 provides the ratings for teachers' performance on the five indicators measured under the Discourse factor. The six teachers showed overall improvement on the "Complexity of Questions" and "Questioning Ecology" indicators, moving from the Developing Inquiry category level to the Proficient Inquiry level. These suggest that teachers had increased success in engaging students in more open-ended discussions and did a better job challenging students to explain and justify their answers.

The "Communication Pattern" indicator did not show any overall significant gains from the Developing Inquiry pre-Institute levels. This indicates that teachers mostly relied on didactic pattern of communication and directed most of the conversation. Providing more resources for teachers that show examples of effective ways of getting student-generated questions to drive the lesson could help improve this factor.

Across all teachers' ratings for Discourse constructs, Pre-Inquiry was selected five times, occurring three times for one teacher (a second observation of this teacher showed notable increases on this construct). The Exemplary Inquiry rating was only selected one time, occurring in the "Communication Pattern" factor. The overall ratings for Discourse showed teachers to have achieved a level of Proficient for the lessons observed.

Table 3. Comparison of Cohort 1 observations scores for Discourse Factors

		Questioning Level	Complexity of Questions	Questioning Ecology	Communication Pattern	Classroom Interactions
Pre-Institute	Mode	3	2	1	2	2
	Median	3	2	1.5	2	2
Post-Institute	Mode	3	3	3	2	3
	Median	3	3	3	2	2

Assessment

The ratings for teachers' performance on the five indicators measured under the Assessment factor are shown in Table 4. Overall teachers performed at a Developing Inquiry level regarding how they assessed prior knowledge. This indicates that while teachers did assess students' prior knowledge, there was no evidence that they used this to modify their instruction. It could be possible that knowledge was used to change the delivery of the larger "unit" plan, but that could not be captured in a cross-sectional observation.

Teachers showed increases in the "Student Reflection" indicator which means they explicitly encouraged students to reflect on their learning at an understanding or authentic level. Additionally, the teacher moved from Developing to Proficient category on the "Role of Assessing" indicator, noting that teachers did solicit information to assess student understanding and took the next step of using that information to adjust the instruction.

Across all teachers' ratings for the Assessment factor, Pre-Inquiry was selected five times, but four of those occurring exclusively for one teacher (the same observation discussed in the Discourse section above. The subsequent evaluation showed increases across all indicators in this factor as well, including Exemplary ratings for three out of five indicators). The Exemplary Inquiry rating was selected twelve times overall (compared to only twice in pre-Institute), with one teacher receiving Exemplary rating on all five factors.

Table 4. Comparison of Cohort 1 observations scores for Assessment Factors

		Prior Knowledge	Conceptual Development	Student Reflection	Assessment Type	Role of Assessing
Pre-Institute	Mode	2, 3	3	2, 3	2	2
	Median	2.5	3	2.5	2.5	2
Post-Institute	Mode	2	3	4	2	3
	Median	2	3	3	3	3

Curriculum

Table 5 reports the ratings for teachers' performance on the four indicators measured under the Curriculum factor. Just as with the pre-Institute observations, the six teachers scored lowest on the "Organizing and Recording Information" indicator, performing just above the Developing Inquiry level of proficiency. These ratings indicate that students mostly recorded information in prescriptive ways with minor input regarding how to organize and record the information.

Teachers showed some gains on the "Learner Centrality" indicator, moving from the Developing Inquiry to Proficient Inquiry rating overall. This shift resulted from teachers moving away from reliance on prescribed activities with anticipated results and going toward more flexibility during the investigation that allowed for student designed exploration.

Across all teachers' ratings for Curriculum constructs, Pre-Inquiry was selected just once compared to three times it occurred in pre-Institute observations. The Exemplary Inquiry rating was indicated ten times, while it was not selected at all in pre-Institute observations. The Exemplary rating was observed most frequently in the "Integration of Content and Lesson" indicator, suggesting the teachers integrated the content of the lesson with the student investigation seamlessly.

Overall comparison of pre- to post-Institute observations

Prior to participation in the INITIATE Summer Institute, teachers in general displayed inquiry-based instruction at the Developing Inquiry level with the Instructional factor showing the greatest degree of proficiency. When the teachers were observed again after their participation in the Summer Institute there were more occurrences of Exemplary ratings across each of the four factors. Overall, the teachers performed better on the Instructional and Assessment factors than on the Discourse and Assessment factors, but showed evidence of small to moderate gains across all factors. Of the three teachers who were observed twice, two of them showed relatively consistent ratings between the observations. One teacher had a seemingly outlier, low performance on the first observed lesson. Since no qualitative interview was conducted afterward, there can only be speculation as to what might have been the cause. The second observation of this teacher showed notable increases in her performance across all factors, and aligned more closely with her other measures (CT score; level of concern).

Table 5. Comparison of Cohort 1 observations scores for Curriculum Factors

		Content Depth	Learner Centrality	Content - Investigation Integration	Organizing and Recording Information
Pre-Institute	Mode	3	2	3	3
	Median	3	2	3	2.5
Post-Institute	Mode	3	3	4	2, 3
	Median	3	3	3	3

4) Teacher Confidence in Adopting the INITIATE Model (Cohort 1): Evaluation questions 2, 3, 4, 5, & 8

The six teachers participating in the 2018 INITIATE Summer Institute took the Stages of Concern Questionnaire (SoCQ) (http://www.sedl.org/pubs/catalog/items/cbam21.html) prior to the onset of the Institute, in fall 2018, and again in February 2019. The assessment measures the level of concern the teachers may have regarding implementing or facilitating innovation (change) in their schools or classrooms. The theory is that the more comfortable a teacher is with the task of implementing the innovation, the more likely it will be introduced to the classroom with fidelity. SoCQ is divided into three major constructs: concern about impact, concern about the task of implementing (logistics), and concern about self (self-efficacy). Respondents are given a series of statements and are asked, using a 7-point scale to indicate their level of agreement with the statement. Anchors within the scale are:

7 = true most of the time

4 = true some of the time

1 = not true at all at this time

0 = this statement is not relevant to me

A score of 0 indicates that the innovation is not a high priority to the respondent. There are six stages of concern and they are illustrated in Figure 1 (note: the lowest step—awareness—was not included in the SoCQ). The stages are developmental in that one progresses from the lowest "step" to the highest.

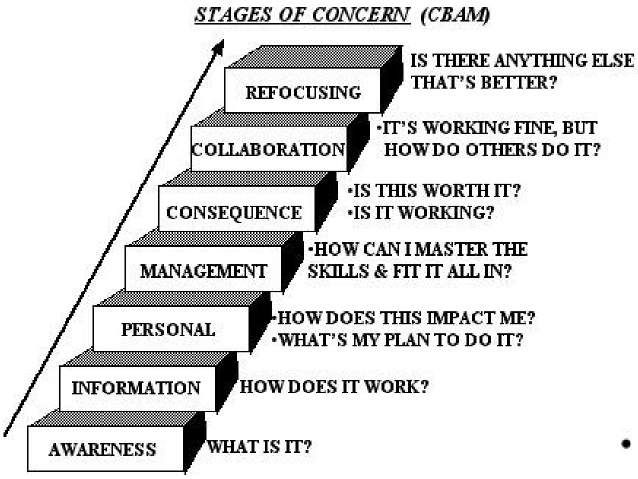

STAGES OF CONCERN (CBAM)
REFOCUSING
IS THERE ANYTHING ELSE THAT'S BETTER?
COLLABORATION
•IT'S WORKING FINE, BUT HOW DO OTHERS DO IT?
CONSEQUENCE
•IS THIS WORTH IT?
•IS IT WORKING?
MANAGEMENT
•HOW CAN I MASTER THE SKILLS & FIT IT ALL IN?
PERSONAL
•HOW DOES THIS IMPACT ME?
•WHAT'S MY PLAN TO DO IT?
INFORMATION
HOW DOES IT WORK?
AWARENESS
WHAT IS IT?

Figure 1: Six stages of concern

Figure 2 illustrates the three administration percentile comparisons for each of the six stages plus 0 (not a priority). Figure 2 provides the group percentile scores for each testing period and shows that the pre Institute testing (blue on Figure 2 chart) indicated that the teachers' scores fell primarily between *unconcerned* and *informational* meaning that they were interested in the INITIATE model but were not quite sure it was relevant to their teaching and needed more information regarding the specifics of the innovation before they would be willing to implement it. The spike for *management* (40%) indicated that some of the teachers had concerns about how to implement the innovation with regards to organizing, scheduling, and efficiency.

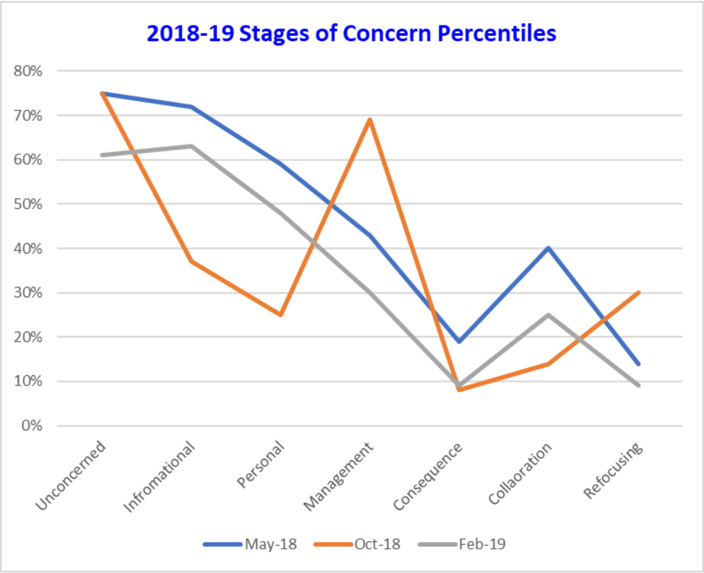


Figure 2: Overall stages of concern over time

Post participation testing in October (orange line), roughly four months after participation and 2.5 months after 2018-19 school year commenced, showed that teachers continued to predominantly be *unconcerned* suggesting that they were still not certain how the INITIATE model would fit into their teaching. However, the fact that the *management* level was nearly as high as *unconcerned* suggests that they understood enough about the model to begin to ponder exactly how it might be scheduled and managed in their classrooms. During this period 30% of the teachers also had concerns about *refocusing.* This indicates that the teachers had other ideas that carried more weight than this innovation at that time.

The second posttest (gray line, February 2019) showed a slight decrease in the percent who were, in general, *unconcerned* (61%). Responses during this testing mirrored the pretest scores; however, in every stage the teachers scored lower on the scales than they did on the pretest. The *informational* stage did see a slight increase suggesting that teachers had an interest in the innovation but are not committed to implementing it. In fact, this level typically indicates a general interest pending more information.

Overall the teachers appeared, over time, to lack a clear understanding of the INITIATE teaching model and how it can be implemented into the classroom. Moreover, the lowest scoring category at each testing was *consequence* or focus on the impact on students. Teachers had little concern for how to evaluate student outcomes and how they may be linked to competencies. This category also focuses on teachers' concerns about changes that may be needed to be made to improve student outcomes. That this category was consistently low suggests that teachers have not moved past concern about implementing the model from their own perspective and started to consider ways in which the model will improve student learning.

Individual findings: Six teachers comprised Cohort 1 (2018-19). Table 6 contains individual percentiles for each stage at each testing. Stage numbers represent the following stages:

0 = unconcerned 1 = Informational 2 = Personal 3 = Management
4 = Consequence 5 = Collaboration 6 = Refocusing

Table 6. Individual stages of concern progress

ID	Date	0	1	2	3	4	5	6
001	5/18	40%	98%	72%	52%	38%	52%	17%
	10/18	Did not participate in this testing						
	2/19	40%	27%	21%	52%	16%	31%	14%
003	5/18	94%	75%	67%	30%	16%	22%	9%
	10/18	97%	45%	31%	23%	7%	16%	9%
	2/19	87%	30%	25%	73%	8%	12%	30%
004	5/18	14%	27%	14%	7%	9%	55%	6%
	10/18	22%	34%	12%	43%	3%	19%	5%
	2/19	69%	30%	48%	65%	16%	22%	34%
005	5/18	91%	69%	70%	77%	19%	59%	11%
	10/18	55%	37%	52%	77%	16%	88%	11%
	2/19	98%	60%	21%	97%	16%	25%	65%
006	5/18	75%	60%	45%	11%	5%	16%	3%
	10/18	48%	40%	35%	27%	8%	14%	17%
	2/19	0%	5%	5%	2%	1%	3%	14%
007	5/18	98%	97%	80%	83%	48%	31%	52%
	10/18	87%	57%	63%	69%	30%	22%	34%
	2/19	99%	48%	25%	97%	5%	9%	26%

The highest percentile is highlighted for each teacher at each testing. As recommended by the *Stages of Concern Instrument Manual*, when another stage score is within one or two percentile points of the highest score, both scores have been highlighted. Concerning the adoption of an innovation, the typical non-user profile will have high scores for Stages 0 – 2 and low scores for 4 -6. Looking at the pretest scores, all teachers except 004 scored as non-users. The cohort was comprised of five math teachers and one CTE teacher. Teacher 004 was the CTE teacher and most likely entered the project with pre-conceived ideas about what the INITIATE professional development would entail.

The typical user will have the highest score at Stage 3 or above. At the October posttest, teachers 004 and 005 were considered innovation users. Teacher 001 did not complete the first posttest and the other teachers remained in the non-user category. In fact, at that point, none of the teachers had attempted to implement the new instruction into the classroom and may not have considered the details of doing so (unconcerned or more concerned about other things). Teacher 004 was most concerned with managing the implementation within the classroom including time and logistics. Teacher 004's concerns shifted from *collaboration* or an interest in working with his or her colleagues on the pretest to coordinate the use of the innovation (in fact, the purpose of INITIATE) to *management* suggesting that at the posttest the teacher was more concerned about how to personally implement the innovation rather than something more advanced like working together with others towards an integrated implementation. On the other hand, teacher 005 jumped from *unconcerned* on the pretest to an interest in *collaborating* on the first posttest. This teacher was embracing the use of the innovation but, because the lessons had not yet been implemented in the classroom, may not have thought through issues associated with the lower Stages such as implementation details and student outcomes.

The February 2019 posttest reflects effects of implementing the INITIATE lessons on the teachers' concerns. First, according to the *Manual*, a higher score for Stage 6 than for Stages 4 and/or 5 indicates that the respondent has ideas that have more merit than the proposed innovation. Scores highlighted in Stage 6 in light green are such occasions. Note that by the second posttest, all the teachers fell into this category except 001. Teachers 001, 005, and 007 were considered users but it should be noted that teachers 005 and 007 had a multi-modal score between *unconcerned* and *management*. This suggests that while these two teachers were essentially non-users, the fact that they were required by the project to implement the lessons caused them to focus on the logistics of doing so. Of all the teachers, teacher 001 followed what might be considered a typical pattern of moving from a low Stage to a progressively higher one.

Teacher 003 remained at Stage 0, *unconcerned*, consistently. Teacher 004 followed an atypical pattern of embracing the innovation at the onset to an *unconcerned* attitude towards adopting it at the second posttest. The negative shift for teacher 005 between the two posttests was most likely due to implementing the lesson in the interim and, as a result, realizing that details of implementation needed attention. Teacher 006 jumped from *unconcerned* on the pretest and first posttest to *refocusing* on the second posttest (Stage 0 to Stage 6). However, note that all the percentile scores for this teacher are low with 14% as the highest score. This brings into question the authenticity of responses and it is recommended follow up with this teacher to clarify his/her attitude towards the program.

Summary: At this point some of the teachers in Cohort 1 have moved to a user status but at the lowest level. Teachers may not understand the INITIATE innovation or at least how it should fit into their teaching and, by the second posttest, it appears that their attitude in general was that they are non-users or are using the innovation because it is a required component of participation. Further information from the teachers was recommended and as a result a focus group interview was conducted May 2019. Through this interview, a better understanding of resistance to adopting the model was realized. Combined with classroom observations and teacher feedback, a clear relationship between teachers' implementation and their attitude about the INITIATE lessons was revealed. As a result, modifications to the 2019 PD was enacted and are described later in this report.

5) Teacher Focus Group Interviews (Cohort 1): Evaluation questions 4 & 6

June 2018 Interview: Findings from a post-participation focus group interview with teachers last June were provided in the previous evaluation report. To summarize, teachers made a few suggestions for next year: First they felt that sitting in a classroom for "two hours at a pop" was a bit tedious and hoped the segments might be mixed up a bit. They also felt the lunch break (1 hour) was too long and preferred more hands-on time with content. There were aspects of the Institute that were confusing to them and it was not until the conclusion that some made sense. The teachers did provide a list of concerns to the evaluator and they correlate with the findings of the SoCQ results in the previous section:

- How can we fit these lessons into our already tight schedule?
- Will this method really improve mathematics learning?
- If they return next year, how will new teachers be integrated into the project? (concerned that next year may just be a repeat)

At this point, all of the teachers felt confident they could integrate their lessons into their classes in the academic year. One suggestion regarding next year was the possibility that INITIATE run academic year meetings at the school in the spring prior to the Institute to help the teachers become familiar with programming the cars to better prepare them for the Institute.

April 2019 Exit Interview: This interview took place near the conclusion of Cohort 1's participation. Teachers provided feedback on several aspects of INITIATE:

Beginning the Program and Summer Institute: The teachers described the Summer Institute as a great experience overall. They particularly enjoyed the field trip and seeing the autonomous car in person because that provided hands-on learning experiences and information about how the cars worked. They were pleasantly surprised by the Institute as some were not looking forward to the summer component. In hindsight, they thought not all of the time in the Institute was used effectively, mentioning painting cars and drawing pictures. The teachers enjoyed the ability to problem solve but felt they could not recreate that experience for their students due to time constraints in their classrooms. Additionally, they said they were never told about the culminating project during the summer. One of their major complaints was that the focus expectations of the program were not clearly established and eventually caused confusion and disinterest in the teachers.

Implementing the Program: Teachers unanimously found the camaraderie of the academic year meetings to be helpful in implementing the INITIATE program throughout the school year, but found the input from project staff to be confusing or conflicting with their (teachers') established understanding of what they were supposed to be doing. Teachers said they felt preached at and that UT wanted to changed everything they were doing yet UT didn't know the teachers' students and their capabilities. Teachers felt that program staff tried to change driving questions midway through the year and that caused confusion and irritation.

INITIATE Lessons: Teachers implemented the lessons as part of a section of their text book and said they did not know the lessons needed to be connected within the class. Some teachers taught a specific concept (e.g. area of polygons) in a traditional format then used the INITIATE lesson as a verification or an additional activity rather than as the method for teaching the concept.

The most prevalent critique/concern expressed by the teachers was the time constraints. They felt that a 47-minute class was not sufficient to teach INITIATE lessons and they seemed reluctant to stretch these lessons/approaches out over multiple or successive class periods. The teachers felt they could not get to all of the program components (programming, aspects of smart cars, problem-based learning, computational thinking). Rather than integrating these components into their teaching style, they viewed them as separate topics. They also felt that there were too many components to be able to incorporate successfully. The teachers felt that the programming aspect – while a useful skill – was too far outside of the scope of what they needed to cover.

Teachers found that some students responded better to these lessons and others just went through the motions. Teachers frequently used worksheets with prescriptive or expected answers as method of assessing student learning. Rather than facilitate real-world discussion, one teacher found the outcomes were almost never reached (due to tech issues mostly) and that was inhibitive. As a result, several teachers felt that the lessons did not help the students learn computational thinking concepts.

While the INITIATE lessons did facilitate a change in teaching style, this change was more a result of having to work more one-on-one with students to solve tech issues. In fact, teachers complained frequently about tech issues: programming taking too much time, tablets not responding, and cars not being calibrated. A lot of these issues discouraged teachers from using the smart car aspect, even though they liked other components of the lessons and pedagogical approach. Additionally, teachers felt they could not spend as much time on INITIATE style lessons because the content that they developed lessons for were not huge components of the state standardized test.

Teacher perceptions of INITIATE expectations: Overall, teachers felt that INITIATE tried to do too much at once. Compounding the difficulty, expectations were not always clearly defined. Teachers felt like overloaded and a potential downfall for any subsequent group. Teachers felt program staff were not clear on the actual focus of program amongst the various components.

This caused at least one teacher to treat the program as something to "check off the to-do-list," and several others commented on their disengagement over time. As a result, the program began to feel more "intrusive" rather than a chance to explore and experiment with different teaching strategies.

Culminating project: Teachers were not averse to the idea of a culminating project but felt it needed to evolve from their input. They felt that the program would be more successful in phases, with teachers learning about and introducing smart cars early, then later on developing an integrated project. Teachers did not know about a culminating project until too late and that it would need to be very clear how the lessons would be expected to be integrated not only within a single teacher's classroom, but across all the classrooms.

6) Teacher Computational Thinking Assessment: Evaluation question 1

Teachers completed the Computational Thinking assessment (CT) on the first and last days of the summer institute, and again 8 months after the summer institute. The Teacher CT assessment was designed by project staff and evaluators to gauge participant knowledge and understanding of computational thinking concepts and how to implement them in the classroom. Validity studies for the instrument are ongoing as more data are collected and will be reported in a future publication. The current report will discuss: 1) findings from an initial inter-rater reliability study; 2) preliminary validity of the instrument and reliability of the measures of teacher CT knowledge/classroom implementation; and 3) findings from Cohort 1.

Inter-rater reliability

The Teacher CT assessment consists of 13 items. The first four items ask teachers to define a CT concept (algorithm design, pattern recognition, abstraction, and decomposition) and provide something a student might do to exhibit understanding of that concept. Responses are scored on a 4-point scale that rewards accurate definition of concept that clearly links with an appropriate student behavior. The next six items deal with a fictitious student homework assignment where the respondent needs to identify and justify the CT concept involved with different aspects of the assignment. These responses as dichotomized into correct/incorrect identification of the element and sufficient/insufficient justification. The final three questions require respondents to detail how they would integrate CT into specific and general math and CTE classroom environments. Again, answers are scored on a 4-point scale that rewards respondents for depth, clarity, and comprehensiveness of their response. Thus, respondents could score a 0 (very little to no CT knowledge) up to 27 (very high level of CT knowledge).

Including multiple raters is necessary to make sure an estimate of a teacher's latent CT knowledge best approximates their "actual" knowledge, and is not the product of one rater's bias or inexperience. There were seven initial raters tasked with providing ratings to Cohort 1's pre- and first post-SI assessments. To increase the reliability of the assessment of teacher CT knowledge, Krippendorff's alpha was computed. This reliability coefficient measures the agreement among raters. Table 7 reveals that the full group of seven raters nor either of the specific sub-group of raters by organization could meet the recommended level of agreement (i.e. greater than .8). In addition to having a greater alpha level, the three raters identified as having greatest reliability also have a tighter confidence band around that coefficient, indicating a 93% probability of meeting the .8 agreement threshold.

Table 7. Measures of inter-rater reliability for scoring the CT assessment

	Krippendorff	95% Confidence Interval		Prob. of achieving
	alpha	Lower Bound	Upper Bound	> .8 alpha
All raters	0.71	0.65	0.76	< .001
Acumen raters	0.69	0.63	0.74	< .001
UT raters	0.72	0.67	0.77	< .001
Selected raters	0.83	0.79	0.87	.93

There did not appear to be any systematic bias between any given rater and any given item. This suggests that discrepancies in scores assigned to an item were the result of different conceptions of that CT concept. Variances in rubric usage were addressed through a meeting that entailed group deliberation, explanation of the concept in question, and discussions of areas of discrepancy (i.e., "What does an answer to Item 4, 'Abstraction concept and behavior', need to look like to get a score of 3, accurate?)

Preliminary validity assessment

The ratings given to each item are ordinal in nature: responses to each item are scored based on the 2- or 4-point rating scales above. These merely make cut points based on the quality of the response. The resulting raw CT scores do not meet the requirement of being on an interval scale necessary for many statistical analyses. Therefore, the data were analyzed using the Rasch measurement model (RMM). The RMM logarithmically converts ordinal raw scores onto a linear scale. This process also provides important pieces of validity evidence pertaining to how well the instrument meets the fundamentals of measurement: separate people into distinct categories of less or more of the trait; only measure one thing; and the reliability of the items in defining the construct of CT knowledge.

The initial findings indicated potential promise for the CT assessment. The equivalent to the Cronbach alpha was .84, indicating a moderately high global level of reliability. The estimates of teacher CT knowledge level and item difficulty alone could explain 53% of the variance in the responses. Ongoing analyses will continue to explore how well the items fit onto/tap into the theoretically defined developmental map of what it means to have less-to-more CT knowledge/application in the classroom.

Assessment of teacher CT

The Rasch interval measures of teacher CT were used to compare teacher performance on the assessment over time. Table 8 shows the average measure for the six teachers before entering the

INITIATE program was just over 12 while in the assessment periods following their participation in the summer institute their scores went up just over 3 points (25% increase).

Table 8. Average CT measures for Cohort 1

	Mean	N	Std. Deviation	Std. Error Mean
Pre-SI	12.2	6	3.3	1.3
Post-SI_1	15.6	6	3.0	1.2
Post-SI_2	15.8	6	3.4	1.4

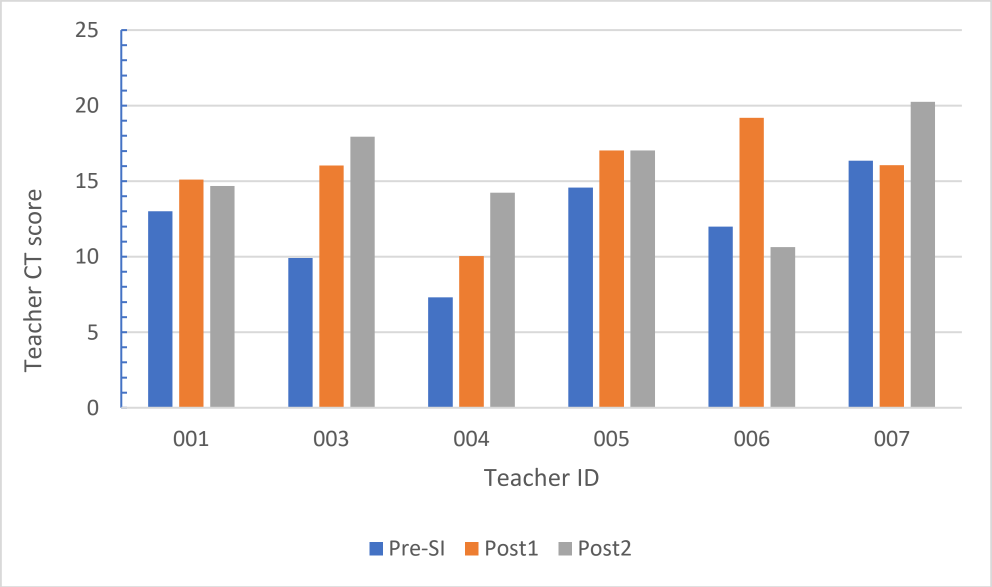
Table 9 shows that the increase Cohort 1 displayed between the start and the end of the two-week long summer institute were statistically significant. Additionally, those gains maintained over the eight months following the summer institute. Further, Hedge's g of .99 for both pairs indicates a large effect size; namely, the pre-SI CT scores are one standard deviation lower than both of the post-SI scores.

Table 9. Comparison of Cohort 1 CT knowledge over time

	Paired Differences								
		Std.	Std. Error	95% Confidence				Sig. (2-	
	Mean	Deviation	Mean	Lower	Upper	t	df	tailed)	Hedge's g
Pre_SI - Post_SI_1	-3.38333	2.77375	1.13238	-6.29420	-0.47247	-2.988	5	0.03	0.991
Pre_SI - Post_SI_2	-3.58333	3.47529	1.41878	-7.23043	0.06376	-2.526	5	0.05	0.997

Beyond the increase seen at the Cohort level overall, an evaluation of each individual teacher's path along the CT continuum revealed several important findings. The trends in Figure 3 show an initial gain in measure of CT knowledge for five of the six teachers following completion of the summer institute. Teacher 007 was well within the standard error of the measure, moving down less than one-third of a point. The growth of teachers 003 (6 points, or 60%) and 006 (just

Figure 3. Comparison of Cohort 1, by individual, CT knowledge over time



over 7 points, or 58.3%) point to significant gains. Teacher 006 did not sustain the gains but triangulating data from other measures and feedback in the interview, this teacher may have experienced project-fatigue and did not engage with the instrument faithfully (e.g. completed the assessment in 14 minutes, while the average completion time was closer to 30 minutes across all testing occasions to date). Teacher 003 continued to show some growth in CT knowledge/classroom practice over time, while teacher 004, the CTE teacher, showed significant improvement in the later assessment by nearly doubling their score.

7) Summer Institute Observations: Evaluation question 6

Changes were made to the 2019 Summer Institute based upon focus group interviews with Cohort 1 teachers. As a result, the Institute was more organized and offered more hands-on opportunities for teachers to "play" with the concepts and how they could be integrated into their classrooms. Acumen staff observed most of the sessions and scored them using the Horizon Professional Development Observation rubric. The data has been collected. It will be analyzed and shared with INITIATE leadership later this summer and reported in the next evaluation report.

8) Student outcomes: Evaluation question 5

Several instruments were administered to Cohort 1 students in spring 2019. These included a demographic survey to verify student group equivalency and to provide variables that will allow the research team to better drill down their analyses, a student engagement survey, and a student computational thinking survey. The engagement survey, developed through collaboration between INITIATE leadership and Acumen, included three subscales for engagement: Behavioral (observable behaviors students carry out that are necessary to academic success), affective (emotions in the learning process), and cognitive (student perseverance and use of cognitive strategies) (based upon Fung, Tan, & Chen, 2018 and Shernoff , 2001). The computational thinking assessment was also developed through collaboration between INITIATE and Acumen and included items that tested student ability to employ computational thinking. Data is currently being analyzed and results will be shared with INITIATE leadership late summer as well as included in the next evaluation report.

9) Conclusions

Procedural suggestions recommended by the teachers were reviewed by INITIATE leadership and, as a result, changes were made to the 2019 Summer Institute and the academic year follow up. Anecdotal comments from Acumen observers as well as INITIATE staff and teacher participants imply that the changes have improved the program. Data-referenced findings will be calculated throughout the next year.

Table 10 compares evaluation findings to date with the eight evaluation questions. To date the INITIATE project is enacted as designed and on schedule. Enrollment in 2019 has increased and modifications have been made to the PD to provide teachers with more time, more relevant content, and more time to practice creating lessons that reflect Institute intentions into their classes.

Table 10. Evaluation questions

1. Does INITIATE teacher PD improve teacher understanding of CT standards and PBL?	Early evidence suggests that Cohort 1 showed significant improvement in CT knowledge as measured by project-designed CT assessment.
2. Do participating teachers implement the treatment protocol (i.e., PBL and inquiry-based instruction) with fidelity?	The increase in median scores on several indicators of the EQUIP and greater frequency of Exemplary ratings on the inquiry factors provide evidence of teachers increased implementation of PBL in the classroom.
3. Do teachers integrate CT into their mathematics and computer science teaching?	EQUIP findings support a positive integration of CT into teaching. This will be triangulated with student CT test next year.
4. Does lesson sharing among mathematics teachers result in lessons that address the intentions of INITIATE? Do teachers use the lessons and are they implemented as designed?	Cohort 1 did not implement the lessons as designed but did introduce elements of the lessons to their students. Modifications have been made for Cohort 2 to improve communication of project expectations as well as to assist teachers in improving fidelity of treatment protocol.
5. Given #1, #2, and #3 are achieved to satisfactory levels, does this indeed improve student CT, student interest in mathematics, and student mathematics ability?	Student data has been collected but not yet analyzed.
6. Are all project activities offered and implemented as designed? If not, are there justifiable reasons for changes?	Project activities were offered and implemented as designed. Based upon Cohort 1 feedback and INITIATE staff reflection, modifications to improve the program have been incorporated.
7. Based upon findings from answering evaluation questions 1 – 5, does the INITIATE model work?	In year 1 the model worked with some weaknesses. Formative evaluation data has informed changes that appear to improve the effectiveness of the model.
8. Does the Stages of Concern Questionnaire function as an effective tool for diagnosing and addressing teacher concerns and implementation?	Based upon Cohort 1 survey data and triangulated with observations and interview feedback, the instrument appears to diagnose teacher concerns. Data collected from Cohort 2 will verify.

10) References

Fung, F., Tan, C. Y., & Chen, G. (2018). Student engagement and mathematics achievement: Unraveling main and interactive effects. Psychology in the Schools, 55(7), 815-831.Marshall, J. C., & Horton, R. M. (2011). The relationship of teacher-facilitated, inquiry-based instruction to student higher-order thinking. School Science and Mathematics, 111(3), 93-101.

Shernoff, D. J. (2001). The experience of student engagement in high school classrooms: A phenomenological perspective (Doctoral dissertation, University of Chicago, Department of Education).