Knology®

INFACT

INFACT Efficacy Report

December 6, 2022

Bennett Attaway & John Voiklis

Recommended Citation Attaway, B., & Voiklis, J. (2022). *INFACT Efficacy Report*. Knology Publication # EDU.051.602.03. Knology.

Date of Publication **December 6, 2022**

Acknowledgements: INFACT is a collaboration between EdGE at TERC, Digital Promise, Florida State University, FunAtomic, Knology, Looking Glass Ventures / Edfinity, University of Florida, and University of Maryland.

Funding: The contents of this publication were developed under a grant supported by the Education Innovation and Research Program (EIR) of the U.S. Department of Education under the project *INFACT: Include Neurodiversity in Foundational & Applied Computational Thinking,* award U411C190179. However, those contents do not necessarily represent the policy of the Department of Education, and endorsement by the Federal Government should not be assumed.

Knology produces practical social science for a better world.

tel: (347) 766-3399
40 Exchange Pl. Suite 1403
New York, NY 10005

tel: (442) 222-8814
3630 Ocean Ranch Blvd.
Oceanside, CA 92056

Knology Publication # EDU.051.602.03

Table of Contents

Knology Publication # EDU.051.602.03

List of Tables

List of Figures

Introduction

Include Neurodiversity in Foundational and Applied Computational Thinking (INFACT) is a Computational Thinking (CT) platform and collection of materials designed to be used in grades 3-8 with students across a range of cognitive skills and learning needs. In particular, INFACT is intended to help students who struggle with the executive functions of attention, working memory, and information processing. The materials offer a variety of online and offline activities organized into four modules, allowing teachers to choose the options they feel will work best for their class.

The evaluation of INFACT used a cluster quasi-experimental design at the school level (although only one school had multiple teachers enrolled in the study). Teachers in the Treatment condition were expected to implement the INFACT program for at least 10 hours over the course of a semester. The Comparison group was a business-as-usual condition: teachers at the same grade levels spent at least 10 hours teaching CT through other activities or programs (such as Hour of Code). For teachers who worked with multiple classes (such as elementary school math specialists), all classes were in the same condition (Treatment or Comparison).

Elementary- and middle-school teachers across the US who identified their classes as including 20% or more neurodivergent students were recruited for the study, and were non-randomly assigned to the Treatment or Comparison group. (While researchers attempted to keep the groups balanced, not all teachers had capacity to implement a new CT program, and not all had an existing business-as-usual curriculum to use.) Recruitment was done by reaching out to individual teachers, who then obtained permission from their school/district to participate. All students in the classroom participated in CT learning activities, but only those with both parental consent and complete assessment data were included in analysis.

CT practices were assessed at baseline and semester end using a set of puzzles (Interactive Assessment of CT, or IACT) previously designed by TERC to measure CT proficiency of elementary and middle school students and re-validated prior to the INFACT efficacy study. Executive function was assessed at baseline using Adaptive Cognitive Evaluation (developed by Neuroscape), a game-like implementation of standard psychological tests designed to be accessible for a wide range of ages and abilities.

Background

Computational Thinking (CT) is a growing area of STEM education first defined by Wing (2006; see also Papert, 1993, although he did not use this term) as a set of skills that "*involves solving problems, designing systems, and understanding human behavior, by drawing on the concepts fundamental to computer science. Computational thinking includes a range of mental tools that reflect the breadth of the field of computer science*" (Wing, 2006). Multiple groups of researchers (e.g., Barr, Harrison, & Conery, 2011; Brennan &

Resnick, 2012; Weintrop et al., 2016, Shute et al., 2017) have developed frameworks for teaching, assessing, and studying CT, although there is no consensus definition as yet. In fact, a recent review article (Kite et al., 2021) found that approximately half the studies they examined defined CT as a set of code-centric skills while the other half defined it as a set of interdisciplinary practices. The implementation team for INFACT draws from both of these viewpoints to create a set of materials that teaches skills such as problem decomposition, algorithms, and conditional logic, but encourages drawing connections between these skills and situations outside the context of computing.

Although many CT interventions use block-based programming for robots or on platforms such as Scratch to introduce these concepts, CT itself is not inherently tied to computer programming (Li et al., 2020), and researchers have argued that it is best incorporated across core subjects, especially for younger students (Grover & Pea, 2018).

Recent reviews have identified a growing number of studies focused on CT interventions for K-12 students, especially elementary and middle school students (Merino-Almero et al., 2021; Tang et al., 2020; Kite, Park, & Wiebe, 2021; Ezeamuzie & Leung, 2022). These interventions tend to focus on a set of "CT practices," often including problem decomposition, algorithm design, abstraction, debugging, pattern recognition, generalization, and specific programming-related concepts such as the use of conditional logic and loops. Interventions which integrate CT into core curriculum (rather than treating it as a separate topic) were uncommon and typically limited to STEM subjects when they did occur. The majority of interventions, even for young students, focused primarily on programming activities, whether through a blocks-based language such as Scratch or the use of robots such as Bee Bot (programmed by a series of button presses) or KIBO (programmed by inserting physical blocks representing commands into the robot). While the latter are sometimes described as "unplugged" (Bati, 2022), they still require access to specialized equipment.

Neurodiversity is a framework which treats individual differences in thinking and learning as part of normal human variation, rather than the deficit-based model of "learning disabilities." The label "neurodivergent" encompasses those who differ from the "neurotypical" population that deficit-based models would consider default. Some examples of neurodivergence are dyslexia, ADHD, and autism. Neurodivergent students often need additional support in learning environments designed to a neurotypical standard, so resources for differentiated classroom instruction are part of helping these students reach their full potential. The National Survey of Children's Health (2022) reports that around 8.9% of children aged 3-17 have been identified as ADHD, and 2.9% are on the autism spectrum, meaning millions of students nationwide could benefit from learning materials which better accommodate their needs.

INFACT differs from existing computational thinking interventions in that it is designed for use with neurodiverse groups of students and includes embedded support for executive function. It also includes a wide range of truly "unplugged" activities which do not require access to the Internet or specialized tools, making it suitable for use in a broader range of settings. It provides a full but flexible curriculum which teachers can easily start using and customize for their classes, and it does not require teachers or students to have a background in computer science or computational thinking.

Study Description

Research Questions

Our evaluation focused on two research questions:

RQ1: What is the difference in computational thinking proficiency of students in grades 3-8 who have had one semester of INFACT, compared to equivalent students who have had one semester of business-as-usual computational thinking activities?

RQ2: To what extent does the INFACT program moderate the effects of individual differences in executive function on external CT assessments for students in grades 3-8, when compared to equivalent students in the business-as-usual condition?

Intervention Condition

The INFACT materials are available through a web portal (infact.terc.edu) which allows teachers to create "sequences" of lessons chosen from a variety of options. The curriculum is broken down into four modules: Introduction to CT, Clear Commands, Conditional Logic, and Repeat Loops. (Two additional modules, Variables and Functions, will be added to the final product, but were not available during the efficacy study). Each module has a four-part structure: activation, foundational activities, applied activities, and wrap-up (often an assessment), with multiple options for activities. Teacher-facing materials for each activity provide thorough instructions and suggestions for guiding struggling students, and some activities include printable worksheets for students or Google Slides for teachers. Students also have access to the portal, and student-facing materials include Zoombinis games with built-in scaffolding tools, pre- and post-assessments, and module quizzes. (Pre- and post-assessments were required for the efficacy evaluation study, while Zoombinis games and module quizzes were optional.)

Teachers in the Treatment condition did not receive formal professional development, but had access to documents describing use of the portal and curriculum, as well as a member of the TERC team who could answer questions via email. They were free to choose any combination of activities, and were not required to complete all four modules as long as they spent at least 10 hours on INFACT activities over the course of the semester (between administration of the pre- and post- assessments).

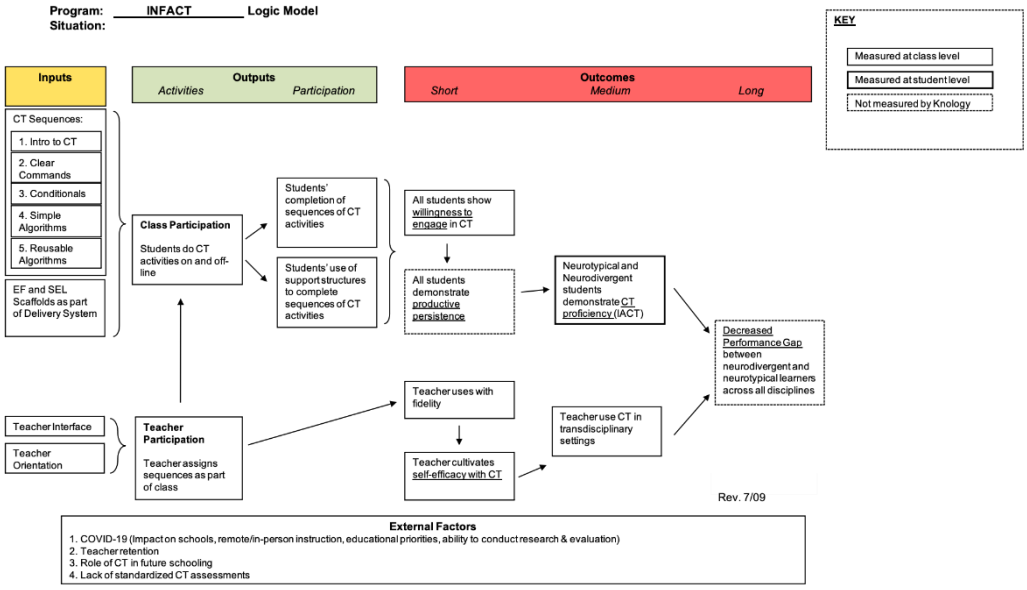

Figure 1. Logic model for INFACT intervention.

Program Implementation

Full participation in the program was defined as at least 10 hours spent on INFACT activities for each class enrolled in the study. The amount of time spent was estimated by teachers during regularly scheduled check-in interviews, which also allowed researchers to ask which activities were chosen and how they were used. In past studies, we have seen low completion rates and levels of detail when requesting teachers to complete a written log of activities. We acknowledged that meeting a 10-hour instructional requirement might not be possible in all cases (for instance, if schools were temporarily closed or many students had to quarantine for COVID-19), given that many teachers in the study were specialists who only saw each class for one period a week. While we required teachers to commit to 10 or more hours of instruction at the beginning of the study, we did not drop them for failing to meet this standard.

Both Treatment and Control teachers varied in role. Some were math or computer specialist teachers who saw students from each grade in the school 1-2 times per week, while others were classroom instructors who worked closely with one group of students (in elementary school) or several sections of the same class (in middle school). However, all teachers administered the intervention or business-as-usual instruction during normal school hours, and none taught it as an elective or selective course. All students in the classroom participated in activities, even if their data was not included in the analytic sample. Both Treatment and Control classes met in person. No activities involved in INFACT posed any risk to students outside what they would encounter in a normal school day.

Data collection was initially planned to take place over a single semester, with 24 Treatment teachers and 24 Control (business-as-usual) teachers. However, this plan was made before the COVID-19 pandemic, which placed increased pressure on teachers even after schools returned to in-person instruction. The actual efficacy study took place over two semesters,

with 6 Treatment and 7 Control teachers participating in Fall 2021 and 8 Treatment and 6 Control teachers (including one Control teacher who also participated in the fall) in Spring 2022. Each teacher enrolled between 1 and 5 classes in the study. Teachers did not receive training as part of the intervention. The majority of participating teachers were women, and while all had taught for several years (many for 10+ years), some Treatment teachers had never specifically taught computational thinking before.

The Fall 2021 cohort participated from October 11 through December 17, and the Spring 2022 cohort from January 24 through May 13 (a slightly longer time period as schools had at least one week of spring vacation). Teachers had 2 weeks in the fall or 3 weeks in the spring to complete pre-assessments with their class, with Treatment teachers gaining access to the INFACT materials only after completing these assessments. Post-assessments had to be completed by the end of the participation period (which coincided with the end of the semester for many teachers). We did not observe notable differences in time spent on INFACT (for Treatment teachers) or business-as-usual CT education between the two semesters.

Given the above considerations, fidelity was defined by two variables, both at the class level:

1. **Class participation** was defined as 75% or more of students engaging in all INFACT activities assigned to them, as reported by teachers.
2. **Teacher participation** was defined as the teacher assigning 5 or more hours of INFACT activities.

At the program level, fidelity was defined as 50% of classrooms meeting the class participation standard and 75% of classrooms meeting the teacher participation standard.

At the teacher level, 23 teachers out of 25 participating in the study met fidelity criteria, for a total of 92.5%. Not all teachers had students included in the analytic sample (see "Study Participants" section below); looking at only teachers who had students included, 91.6% met fidelity criteria.

The teachers in our study all taught in person and actively facilitated the INFACT activities, so they were able to describe accurately whether or not their students had completed an activity. During interviews with teachers, only one mentioned a student choosing not to participate in an activity, and all described that their class as a whole was highly engaged with the content. We can be confident that fidelity criteria for student-level fidelity were met in all classrooms.

Setting

Participating classrooms were all located in the US, but not all in a specific district or state as teachers were recruited individually. We did not ask teachers whether their school was in an urban, suburban, or rural setting, but some mentioned being in a rural area during check-in interviews. Public, parochial, private, and charter schools were all eligible for participation in the study. Among Comparison teachers, nine taught in public schools and three in private

schools. Among Treatment teachers, nine taught in public schools, five in private schools, and one in a charter school.

While there were no requirements for school, location, or setting, we asked all teachers in both conditions to confirm that they typically teach classes including at least 20% neurodivergent students. Specifically, they were asked to *"Confirm that your grade 3-8 classroom population typically includes at least 20-25% neurodiverse [sic] students (e.g., students who have IEP/other classification or teacher/parent designation as needing learning support)."* This approach was preferable to requesting student IEP status for two reasons: (1) to protect individual students' privacy; (2) because IEP status does not correspond directly to neurodivergence. Neurotypical students may have IEPs due to physical disability or injury, and not all neurodivergent students will have an IEP, as obtaining one is an extended process.

Comparison Condition

The Comparison condition was business-as-usual for teachers who committed to teach CT using non-INFACT resources of their choice for at least 10 hours over the course of the semester. These teachers all had at least some prior knowledge of computational thinking (since the study asked them to continue their typical CT instruction). Students would have had access to the CT instruction these teachers provided outside the context of the study.

Each Comparison teacher was free to use any combination of existing CT resources, so the specific activities each conducted with their class varied. However, most classes used Scratch or other coding environments such as Code.org, often in the context of Hour of Code activities. Games or puzzles were also a popular activity, and several classes used robots.

Like the Treatment condition, the Comparison condition was delivered by a range of teachers including subject specialists and classroom teachers. A total of 12 teachers participated over two semesters, with one of these participating in both Fall 2021 and Spring 2022. Teachers were not allowed to participate in the Treatment and Comparison conditions during the same semester, or in the Comparison condition after participating as a Treatment teacher. However, Comparison teachers in Fall 2021 were given the option to participate as Treatment teachers in Spring 2022 if and only if they would be working with a completely new cohort of students. One teacher took advantage of this opportunity. All teachers were welcome to use INFACT materials once their participation in the study was complete.

Study Participants

Participants were invited through advertisements shared in Facebook groups and email lists, and through individual outreach to teachers who had been involved in past CT projects (unrelated to INFACT) run by the implementation team. Teachers were allowed to participate in either the Treatment or Comparison group, although if significantly more teachers were interested in one condition than the other, the implementation team member in charge of recruitment would focus on recruiting for the smaller condition to obtain a more balanced

sample. Additionally, since the business-as-usual condition required teachers to implement a CT curriculum not provided to them by the researchers, recruitment for this group targeted teachers who had been teaching CT prior to the study.

Inclusion criteria were:

- All students enrolled in the study must be at least in 3rd grade and at most in 8th grade.
- Teachers must confirm that they typically work with at least 20% neurodivergent students.

22 Treatment and 24 Control teachers were recruited across the two semesters, of which 15 Treatment and 14 Control teachers enrolled classes in the study. One teacher in each condition dropped out due to timing-related issues, so a total of 14 Treatment and 13 Control teachers completed the study.

However, the analytic sample did not include all of these teachers, as the middle school sample was not within a reasonable threshold of baseline equivalence. Accordingly, we chose to use only elementary school classrooms in our analytic sample, and performed matching to obtain baseline equivalence between conditions (Hedges' $g \leq |\pm 0.03|$). Prior to matching, we removed students with incomplete baseline or outcome assessment data from the sample. The final analytic sample contained five Treatment and seven Control schools, with a total of 182 students per condition.

Table 1. Number of students, teachers, and schools per condition.

School	Condition	Teachers	Students
A	Treatment	1	90
B	Treatment	1	27
C	Treatment	1	5
D	Treatment	1	19
E	Treatment	1	41
Total: 5		**5**	**182**
F	Comparison	2	10
G	Comparison	1	16
H	Comparison	1	53
I	Comparison	1	3
J	Comparison	1	29
K	Comparison	1	17
L	Comparison	1	54
Total: 7		**8**	**182**

The evaluation sample included a non-random sample of the schools, teachers, and students offered the intervention over the course of evaluation. As described above, we recruited middle school (grades 6-8) classes for the study, but found that the students in Control classrooms had much higher scores on the baseline pre-assessment of CT proficiency, to the extent that statistical adjustment would not be appropriate according to

WWC standards. We therefore excluded teachers who taught only middle-school students from the analytic sample (nine Treatment and four Control teachers). Another two Control teachers did not have consenting students who completed both the baseline and post-assessments, and were therefore not part of the analytic sample.

Study Design and Measures

Independence of the Impact Evaluation

The impact evaluation was conducted by Knology independently from the intervention development team. As this study used a quasi-experimental design, random assignment was not conducted. The electronic system which automatically recorded assessment responses was accessible by both Knology and the intervention developers, but these data were read-only. Knology conducted all data cleaning and analysis without input from the intervention developers.

Pre-registration of the Study Design

The study was pre-registered in the Registry of Efficacy and Effectiveness Studies (REES) on October 12, 2021. At this point, although the pre-assessments were available to the first cohort of teachers, no data had yet been collected. The registration number for the project is #7820.1v2.

The research questions registered were:

1. What is the difference in the proficiency in computational thinking of students in grades 3-8 who have one semester of INFACT compared to comparable students who have had one semester of business-as-usual computational thinking activities?

2. To what extent does the INFACT program moderate the effects of individual differences in executive function on external CT assessments for students in grades 3-8, when compared to equivalent students in the business-as-usual condition?

The model specified for Research Question 1 was:

(1) $$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_4 X_{4j} + u_{0j} + u_{1j} X_{3j} + \varepsilon_{ij}$$

Where:

- Y_{ij} is the composite IACT-Adapted score for student i within teacher j
- β_0 is the overall mean composite IACT-Adapted score
- X_{1ij} is the composite Baseline IACT-Adapted score for student i within teacher j
- X_{2ij} is the composite Executive Function (ACE) score for student i within teacher j
- X_{3ij} is the grade level for student i within teacher j

- X_{4j} is the contrast for Treatment vs. Comparison group
- β_1-β_4 are the coefficients for the three covariates (X_1-X_3) and the Treatment contrast (X_4)
- υ_{0j} is the group-level error at level 2, i.e. the random intercept for teacher j
- υ_{1j} is the group-level error at level 2, i.e. the random slope for teacher j
- ε_{ij} is random error at level 1

The model specified for Research Question 2 was:

(2) $$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \beta_3 X_{3ij} + \beta_5 X_{2ij} X_{4j} + u_{0j} + u_{1j} X_{2ij} + \varepsilon_{ij}$$

Where:

- β_5 is the coefficient for the interaction term that models the moderation effect of individual differences in executive function (X_{2ij}) by Treatment (X_{4j})
- All other terms are the same as equation 1

The model specification for Baseline Equivalence was:

(3) $$Y_{1ij} = \beta_0 + \beta_4 X_{4j} + u_{0j} + \varepsilon_{ij}$$

Where:

- Y_{1ij} is the composite Baseline IACT-Adapted score for student i within teacher jβ_0, β_4, X_{4j}, u_{0j}, and ε_{ij} are the same as equation 1

Design

The study used a Quasi-Experimental Design. Neither students nor teachers were matched at the beginning of the study period, as we planned to compare all Treatment students to all Control students. However, we conducted analyses on matched pairs of students once we had student data. We used genetic matching (as implemented in the R package MatchIt) on pre-test score, grade level, and executive function score.

The timing of the study was as follows, and did not differ between conditions:

Table 2. Timing of study milestones in Fall 2021 and Spring 2022 semesters.

Milestone	Fall 2021	Spring 2022
Assignment	At recruitment (prior to October 2021)	At recruitment (prior to mid-January 2022)
Consent	October 2021	January 2022
Baseline Assessment – CT proficiency (IACT) and executive function (ACE)	October 2021	Late January – early February 2022
Intervention Start	Late October 2021	Late February 2022
Intervention End	December 2021	April 2022
Outcome Assessment – CT proficiency (IACT)	December 2021	Late April – early May 2022

Assignment to conditions was done at the teacher level. Once teachers committed to being in the Treatment or Comparison condition, all classes they enrolled in the study became part of that condition. As this was a QED, assignment was not random: teachers chose the condition they participated in. We anticipated that this could cause issues for baseline equivalence of students, and were prepared to drop data for which comparable students in the other condition could not be found.

Measures

Outcome Measure: IACT

The outcome (and baseline) measure for CT proficiency was Interactive Assessments of Computational Thinking (IACT), which had been used in previous large-cohort computational thinking studies by the project implementation team (Rowe et al., 2021). IACT is a game-like assessment designed to require minimal reading and no programming experience from students. It contains four item types targeting the CT skills of Problem Decomposition, Pattern Recognition, Abstraction, and Algorithm Design. It is designed for use with students in grades 3-8. There are elementary school (grades 3-5) and middle school (grades 6-8) versions of the test, containing the same item structure but with differing difficulty levels; results are z-scored to account for this.

We conducted a re-validation of this measure with a small sample of students (n=167) in the summer prior to the efficacy study. This study recruited students in the target age range, with a focus on those whose parents or teachers identified them as neurodivergent. As part of the study, we also had participants complete the ACE measures of executive function and examined the correlation between these and IACT score. A full description of this study is available in the Appendix.

IACT questions take the form of interactive puzzles, which students can experiment with to arrive at an answer. Each module begins with a simple, un-scored task to help students familiarize themselves with the mechanics and goals. None of the four puzzle types corresponds directly to activities in the INFACT curriculum.

Module 1

Items in this module ask students to guess which shape from an array is "correct" through repeated hypothesis testing. After testing an item, students receive feedback on whether each attribute (shape, color, pattern) is correct. Figure 1 shows an easy example problem, where the correct answer is “red diamond.” If a student drags the red circle into the test box, a green check will appear for “color” and a red X will appear for “shape.” A student taking this feedback into account will next try an option which is red, but not a circle, arriving at the correct answer. Scoring is based on the number of moves a student takes to solve the problem compared to the ideal solution.

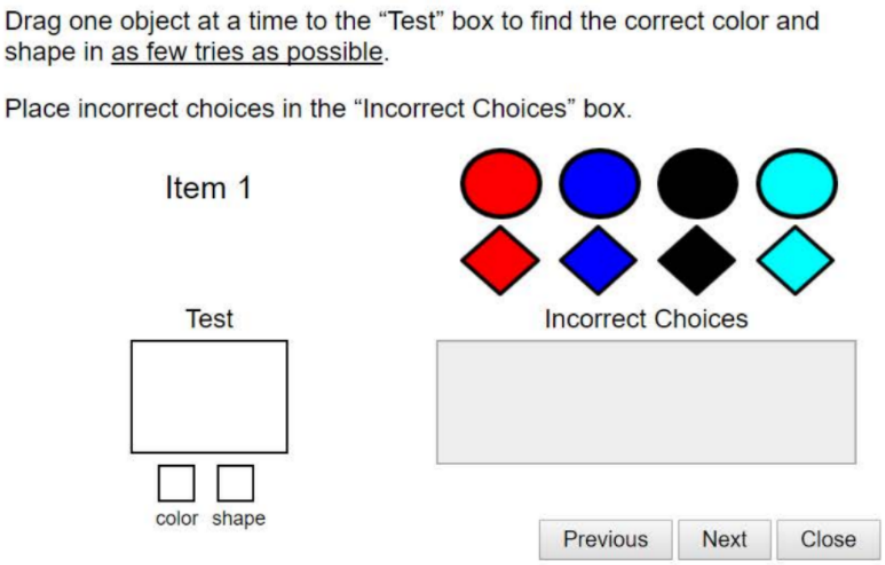

Figure 2. Example item for IACT Module 1 (problem decomposition).

Module 2

This module consists of five items from Raven's Progressive Matrices (Raven, 1981). These items require students to identify the underlying rules being used, and to choose a response which continues the pattern. In the fairly simple example item in Figure 2, the pattern is that in each row, the second and third images are the left and right sections of the first (and that in each column, the second and third images are the top and bottom sections of the first). The correct response is the white square. These items are scored based on whether the answer given is correct, and students are not able to change their answer after clicking Submit.

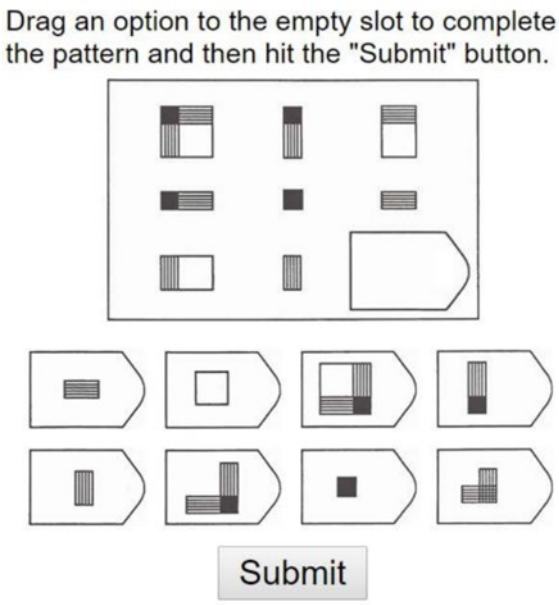

Figure 3. Example item for IACT Module 2 (pattern recognition).

Module 3

This module requires students to infer the underlying rule used to position shapes in a grid. They are presented with a partially-filled grid and asked to fill each of the blanks with the provided shapes to complete the pattern. Each shape can appear only once in the solution. In the example item shown in Figure 3, the pattern is that all items in a row are the same shape and all items in a column are the same color. So, for this example, the solution would be red triangle (top center), blue diamond (middle left), black diamond (middle right), and red circle (bottom center). Later puzzles involve more complex patterns and larger grids. Items in this section are scored as the percentage of empty spaces filled correctly.

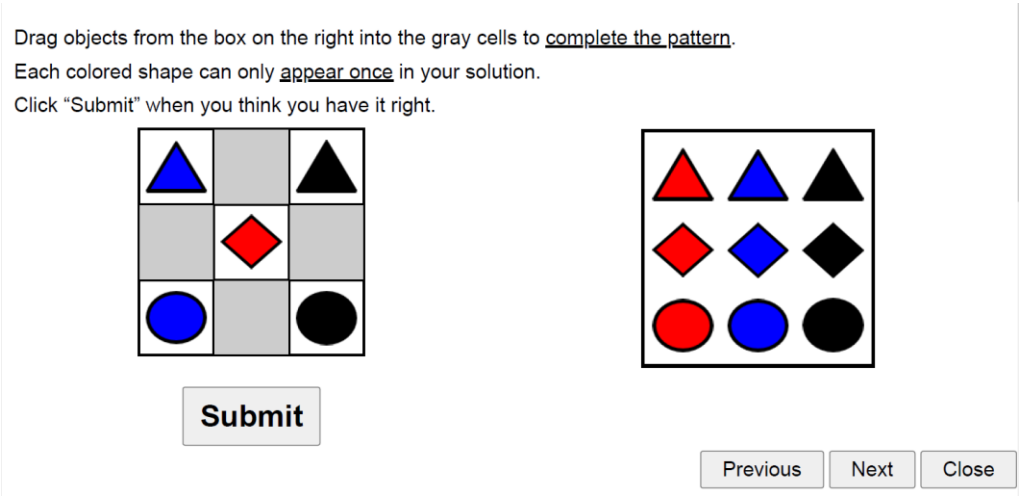

Figure 4. Example item for IACT Module 3 (abstraction).

Module 4

Items in this module involve creating a sequence of arrows that will guide a character along a path, which needs to include several specific points in order to avoid obstacles. In the

example item in Figure 4, the baby panda in the upper left needs to visit the blue, red, and yellow squares in order. Items are scored based on how many moves the student's solution takes compared to the optimal solution.

Some INFACT activities do involve creating a path through a maze, similar to the content of this module, but this type of task is also common in other CT programs such as Code.org / Hour of Code (which many Comparison classrooms used) and robotics tasks. Furthermore, INFACT activities involving path creation do not require students to find the shortest solution, or avoid crossing the same square twice—both of which are necessary for the IACT puzzle.

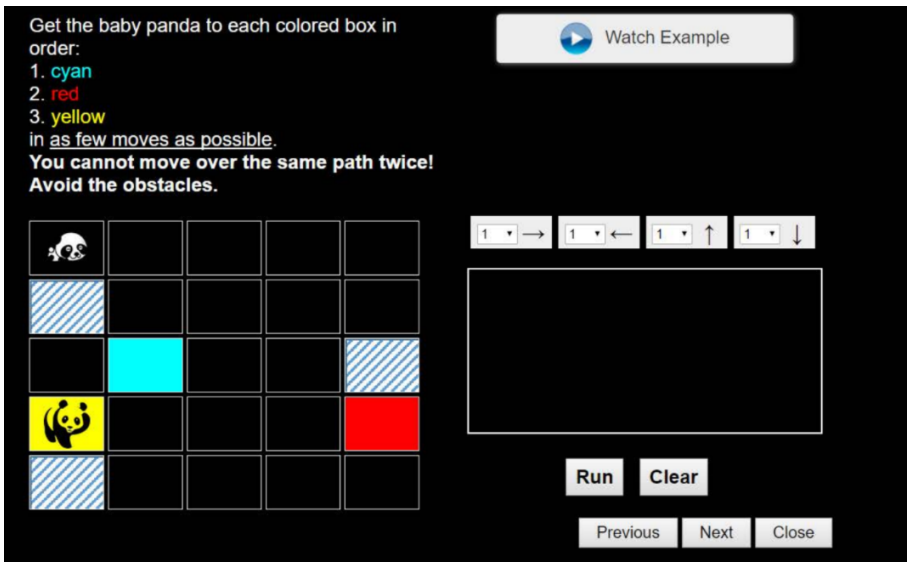

Figure 5. Example item for IACT Module 4 (algorithm design).

Our validation revealed that IACT Module1 showed less variation in scores than the other three, due to students scoring at or close to maximum. Removing this module increased the reliability (Cronbach's α) of the assessment as a whole. Accordingly, we chose to drop this module from the score calculation used in Rowe et al. (2021). This decision was made prior to any data collection for the efficacy study, and was reflected in the pre-registration. The resulting IACT assessment had internal reliability of α = 0.64 on a data set combining the re-validation sample and data from a previous project by the TERC team, and α = 0.74 for the re-validation sample only.

IACT data was collected from both Treatment and Comparison groups immediately prior to the intervention as a baseline measurement, and again at the end of the intervention period to measure outcomes. These two IACT forms ask different questions, so completing the first did not give students an advantage on the second.

Baseline Covariate Measure: ACE

Adaptive Cognitive Evaluation (ACE) is a game-like implementation of standard executive function measures, developed by Neuroscape at UCSF (Coulanges et al., 2021; Mishra et al., 2021). For this study, we used the Flanker, Backwards Spatial Span, Task Switching, and Go/No-Go tasks. We calculated scores for each task using the aceR package developed by Neuroscape and the metrics recommended by the developers for each task (Neuroscape, 2022). For the final analysis, we z-scored each module and averaged the scores across all four to create a single variable. (We experimented with other methods of simplifying the data, including principal components analysis and Mahalanobis distance from a theoretical student with the top score observed for each task, but did not observe significant improvements in model fit.)

Sample Size

Due to technical issues with the IACT assessment in some schools, the number of students with full pre-assessment and post-assessment data was much lower than the number of students who received the intervention.

The number of students with consent for data collection was 373 elementary and 132 middle school students in the Comparison condition, and 260 elementary and 298 middle school students in the Treatment condition.

The number of students with complete baseline data was 278 elementary and 116 middle school students in the Comparison condition, and 222 elementary and 184 middle school students in the Treatment condition.

The number of students with complete data for both baseline and outcome assessments was 257 elementary and 116 middle school students in the Comparison condition, and 190 elementary and 131 middle school students in the Treatment condition. Establishing baseline equivalence further narrowed down the analytic sample, as described in the following section.

Data Analysis and Findings

Baseline Equivalence

We analyzed baseline equivalence at the individual level. In our analysis, we accounted for potential teacher-level effects by including a random effect for Teacher in the model. To test baseline equivalence, we examined Hedges' g for baseline IACT scores. We found that for the 321 Treatment and 373 Control students with full data, *g* was 0.38, indicating lack of baseline

equivalence and falling outside the range in which statistical adjustment is permitted by the WWC.

Upon further inspection, we found that while the elementary school students had comparable baseline scores between conditions, middle school Comparison students strongly outperformed their Treatment counterparts on the baseline assessment. Propensity score matching did not sufficiently reduce this gap. We therefore chose to analyze the effects of INFACT for elementary school students only. After matching Treatment and Comparison students using propensity scores which included baseline IACT, grade level, and executive function score (ACE), we obtained a final analytic sample of 182 Treatment and 182 Comparison students.

Measure	Comparison Group			Treatment Group			Treatment – Control Difference	Standardized Difference
	n	Mean	SD	n	Mean	SD		
IACT (pre)	182	0.697	0.647	182	0.680	0.774	-0.017	**-0.024**
Grade Level	182	4.190	0.799	182	4.088	0.830	-0.102	-0.125
ACE	182	-0.033	0.641	182	-0.165	0.582	-0.132	-0.215

Table 3. Means, standard deviations, and standardized differences for Treatment and Control groups at baseline.

While our preregistered model for baseline equivalence does not require the covariates of grade level and ACE score to be equivalent, we note that the between-group differences for both are low enough to be considered equivalent if a statistical correction is used.

The baseline value for pre-intervention CT skills shows a low enough standardized difference to be considered baseline equivalent with no adjustment. Grade level and ACE values are included in the final model as covariates, as preregistered.

Program Effects

Program-level impacts were measured using a multi-level model which accounted for potential variation due to differences in instruction at the teacher and school levels. Grade level and ACE executive function summary score were included as student-level covariates. Although our analytic sample met the criteria for baseline equivalence without statistical adjustment, we also included baseline IACT score as a covariate, as in the pre-registered model. The analytic model was identical to the pre-registered version except for the addition of a school-level random effect, to account for potential similarities between teachers in the same context. There was only one case in which two teachers worked at the same school. For Research Question 2, we included an interaction term between ACE score and Condition, again matching the pre-registered model, and added the school-level random effect. In both models, students were compared to their matched counterpart in the opposite condition.

Models were fit and effects calculated using the *lmer* and *lmerTest* packages in R. The model for Research Question 1 showed a moderate positive effect on CT proficiency outcomes for students in the Treatment condition (0.41 standard deviations, $p = 0.024$).

The model for Research Question 2 showed a comparable effect size for CT proficiency outcomes of Treatment students (0.39 standard deviations, p = 0.038). We did not observe a statistically significant interaction between ACE executive function score and the Treatment condition, so we cannot conclude whether the intervention was more or less effective depending on student executive function.

While post-hoc power analysis should not be used to argue that a completed study was adequately powered (Zhang et al., 2019), it can be used to argue that the present study was underpowered. Post-hoc analysis indicates that a sample of our size would detect an effect only 55% of the time, if an effect were in fact present.

Table 4. Fixed effects for Research Question 1 Model.

	Estimate	Std. Error	df	t value	p
(Intercept)	0.291	0.149	33.103	1.945	0.060
IACT.pre	0.130	0.075	332.309	1.733	0.084
ACE	0.090	0.094	245.409	0.955	0.340
Grade4	0.129	0.117	281.228	1.107	0.269
Grade5	-0.269	0.120	315.115	-2.233	0.026
Treatment	**0.413**	**0.156**	**10.108**	**2.653**	**0.024**

Table 5. Fixed effects for Research Question 2 Model.

	Estimate	Std. Error	df	t value	p
(Intercept)	0.292	0.152	31.507	1.922	0.064
IACT.pre	0.135	0.750	337.986	1.799	0.073
ACE	0.175	0.116	354.541	1.510	0.132
Grade4	0.119	0.117	285.737	1.017	0.310
Grade5	-0.274	0.120	317.607	-2.274	0.024
Treatment	**0.387**	**0.162**	**10.256**	**2.385**	**0.038**
ACE:Treatment	-0.186	0.151	233.180	-1.233	0.218

We tested a model which added a Semester variable to account for variation between classes participating in the first or second half of the school year, but did not observe improved model fit as measured by AIC or BIC. Because of this, we used the simpler model (without a Semester variable) for analysis. Given that only one teacher participated during both semesters, and was in a different condition for each, it is possible that the random effects calculated for Teacher already capture timing-related variation.

Table 6. Means, standard deviations, and standardized differences between groups on outcome measures.

Measure	Comparison Group			Treatment Group			Treatment – Control Difference	Standardized Difference	p-value
	n	Mean	SD	n	Mean	SD			
IACT.post	182	0.319	0.892	183	0.691	0.760	0.372	0.449	<0.01

Discussion

INFACT is a flexible Computational Thinking (CT) platform and curriculum aimed at students in grades 3-8. Our study addressed two research questions:

RQ1: What is the difference in computational thinking proficiency of students in grades 3-8 who have had one semester of INFACT, compared to equivalent students who have had one semester of business-as-usual computational thinking activities?

In our study, we observed a positive effect on CT proficiency for students in grades 3-5 who received instruction using INFACT as compared to analogous students receiving alternate forms of CT instruction. Due to the underpowered nature of the study, research at a larger scale is needed to confirm this effect.

RQ2: To what extent does the INFACT program moderate the effects of individual differences in executive function on external CT assessments for students in grades 3-8, when compared to equivalent students in the business-as-usual condition?

We were not able to statistically identify an interaction between executive function and INFACT vs. Comparison instruction. As INFACT was designed to include specific supports for students' executive function, this is an area we would like to explore with a more powerful study. We also note that the effect of executive function score on CT performance disappeared post-intervention for students in both conditions, which suggests that computational thinking instruction regardless of curriculum "evens the playing field" for students who score lower on executive function measures.

We note as a caveat that neurodiversity includes a wide range of differences, not only executive function, and that our summary score for executive function cannot capture the full range of variation among neurodivergent students.

Mechanisms and Root Causes

While this report focuses on the quantitative aspect of our study, we also conducted interviews with Treatment teachers and surveys of Comparison teachers, giving us greater insight into mechanisms and root causes. Specifically, these interviews shed light on how neurodivergent students and teachers benefitted from INFACT and how it compared to other CT activities. In particular, we heard that the range of "unplugged" activities available through INFACT were more engaging for high-energy students than computer-based work, and that the visual-heavy materials were valuable for students who struggled with reading. Teachers also reported that the activities were easy to lead, and expressed interest in

incorporating INFACT materials into their instruction moving forward. Engagement is necessary but not sufficient for learning, and this early data suggests one potential route by which INFACT is effective in the inclusion classroom.

The connection between computational thinking and executive function is only beginning to be explored. Arfe et al. (2019) observed increased planning time and response inhibition among neurotypical first-grade students who completed coding lessons, and Robertson et al. (2020) observed a correlation between performance on CT tasks and teacher ratings of executive function. The clearest conceptual overlap is in the areas of goal-setting, planning, and staying on task, which in turn overlap with working memory and attention.

Directions for Future Research & Development

Why and how does INFACT improve the performance of students on measures of computational thinking? Why and how does computational thinking instruction, in general, support the equalization of students with varying executive functioning?

The present study suggests the following directions for future research:

- developing better assessments of computational thinking;
- exploring the relationship between computational thinking and executive functioning through classroom observation; and
- developing and conducting a series of interconnected design studies to identify pathways through which computational thinking in and of itself may serve as scaffolding for executive functioning, and identifying those specific student populations it best supports.

References

Arfe, B., Vardanega, T., Montuori, C., & Lavanga, M. (2019). Coding in primary grades boosts children's executive functions. *Frontiers in Psychology, 10,* 2713.

Barr, D., Harrison, J., & Conery, L. (2011). Computational thinking: A digital age. *Learning & Leading with Technology, March/April*, 20-23.

Bati, K. (2022). A systematic literature review regarding computational thinking and programming in early childhood education. *Education and Information Technologies, 27*(2), 2059–2082. https://doi.org/10.1007/s10639-021-10700-2

Brennan, K., & Resnick, M. (2012, April). New frameworks for studying and assessing the development of computational thinking [Paper presentation]. *The American Education Researcher Association*, Vancouver, Canada.

Child and Adolescent Health Measurement Initiative. (2022). 2020-2021 National Survey of Children's Health (NSCH) data query. Data Resource Center for Child and Adolescent Health supported by the U.S. Department of Health and Human Services, Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB). Retrieved 10/26/2022 from www.childhealthdata.org.

Coulanges, L., Abreu-Mendoza, R. A., Varma, S., Uncapher, M. R., Gazzaley, A., Anguera, J., & Rosenberg-Lee, M. (2021). Linking inhibitory control to math achievement via comparison of conflicting decimal numbers. *Cognition, 214*, 104767. https://doi.org/10.1016/J.COGNITION.2021.104767

Ezeamuzie, N. O., & Leung, J. S. C. (2022). Computational thinking through an empirical lens: A systematic review of literature. *Journal of Educational Computing Research, 60*(2), 481–511. https://doi.org/10.1177/07356331211033158

Grover, S., & Pea, R. (2018). Computational thinking: A competency whose time has come. *Computer Science Education: Perspectives on Teaching and Learning.*

Kite, V., Park, S., & Wiebe, E. (2021). The code-centric nature of computational thinking education: A review of trends and issues in computational thinking education research. *SAGE Open.* https://doi.org/10.1177/21582440211016418

Li, Y., Schoenfeld, A.H., diSessa, A.A., Graesser, A.C., Benson, L.C., English, L.D., & Duschl, R.A. (2020). Computational thinking is more about thinking than computing. *Journal for STEM Education Research, 3,* 1-18.

Merino-Almero, J.M., Gonzales-Calero, J.A., & Cozar-Gutierrez, R. (2022). Computational thinking in K-12 education. An insight through meta-analysis. *Journal of Research on Technology in Education, 54*(3), 410-437.

Mishra, J., Lowenstein, M., Campusano, R., Hu, Y., Diaz-Delgado, J., Ayyoub, Jain, R., & Gazzaley, A. (2021). Closed-loop neurofeedback of α synchrony during goal-directed attention. *Journal of Neuroscience, 41*(26), 5699–5710. https://doi.org/10.1523/JNEUROSCI.3235-20.2021

Neuroscape (2022). *ACE Analytics*. Retrieved October 26, 2022 from https://neuroscape.ucsf.edu/ace-analytics/.

Papert, S. A. (1993). *Mindstorms: Children, Computers, And Powerful Ideas* (2nd ed). Basic Books.

Robertson, J., Gray, S., Toye, M., & Booth, J. (2020). The relationship between executive functions and computational thinking. *International Journal of Computer Science Education in Schools, 3*(4).

Raven, J. C. (1981). Manual for Raven's progressive matrices and vocabulary scales. Research supplement No.1: The 1979 British standardisation of the standard progressive matrices and mill hill vocabulary scales, together with comparative data from earlier studies in the UK, US, Canada, Germany and Ireland. San Antonio, TX: Harcourt Assessment.

Rowe, E., Asbell-Clarke, J., Almeda, M. V., Gasca, S., Edwards, T., Bardar, E., Shute, V., & Ventura, M. (2021). Interactive Assessments of CT (IACT): Digital interactive logic puzzles to assess computational thinking in grades 3–8. *International Journal of Computer Science Education in Schools, 5*(2), 28–73. https://doi.org/10.21585/ijcses.v5i1.149

Shute, V. J., Sun, C., & Asbell-Clarke, J. (2017). Demystifying computational thinking. *Educational Research Review, 22*, 142–158. https://doi.org/10.1016/j.edurev.2017.09.003

Tang, X., Yin, Y., Lin, Q., Hadad, R., & Zhai, X. (2020). Assessing computational thinking: A systematic review of empirical studies. *Computers & Education, 148*.

Weintrop, D., Beheshti, E., Horn, M., Orton, K., Jona, K., Trouille, L., & Wilensky, U. (2016). Defining computational thinking for mathematics and science classrooms. *Journal of Science Education and Technology, 25*(1), 127–147. https://doi.org/10.1007/s10956-015-9581-5

Wing, J. (2006). Computational thinking. *Proceedings of the ACM, 49*(3), 33-35.

Appendix: Re-Validation of IACT

Internal Reliability of IACT

We examined IACT scores from two sources: data collected prior to the INFACT efficacy study (N = 167, ACE scores collected for Comparison) and pre-test data from CodePlay (NSF Award #1738574), a previous collaboration between TERC and Knology which used IACT as an assessment measure (N = 3699, around 25% of students had active IEPs).

The IACT assessment has four modules, each using a different type of logic puzzle. Scoring is done by calculating a raw score for each module (depending on the puzzle type, this could be ratio of moves taken to minimum moves required or percent correct answers), then converting these raw scores to Z-scores based on the data for the full sample. We combined both data sets for these calculations.

As each set of puzzles has a time limit, an NA score for any given model typically means that a student did not interact with that puzzle at all. Lack of interaction with the activity could be due to distraction or noncompliance, not necessarily an unsuccessful attempt to solve the puzzle (especially given that any interaction with the puzzle, even if an answer was not submitted, would be recorded).

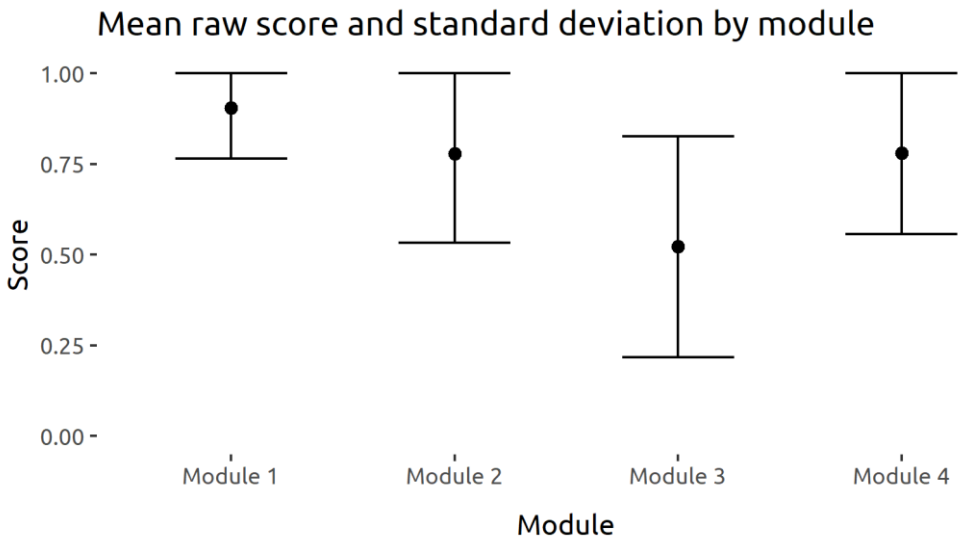

Figure 6. Raw scores and standard deviations on IACT modules 1-4, validation study.

Examining the distribution of scores (Figure 6), we see that Module 1 has notably less variation in scores than the other three modules. In fact, 42% of respondents obtained the maximum score on Module 1, and 72% scored at least 0.9. This skew means that when the module is Z-scored, low scores are heavily penalized.

Although there are separate assessment forms for elementary (grades 3-5) and middle school (grades 6-8), the only difference is that puzzles in the middle-school form are slightly more difficult (for instance, a larger set of answer options or more spaces to fill in a grid). The skills being tested are identical, so we do not separate analysis into elementary and middle school subsets.

Reliability

Initial reliability testing for the combined data set gave a Cohen's alpha score of only 0.52. However, this increases to 0.64 if the Z-score for Module 1 is excluded from analysis. Dropping any other module's score results in a lower alpha. These results are shown in Table 7.

Table 7. Cohen's alpha score when a module is dropped (combined data set)

Module Dropped	*n*	alpha
none	6470	0.51
1	**6470**	**0.64**
2	6470	0.41
3	6470	0.37
4	6470	0.30

Looking only at the data newly collected for this validity study (Table 8), Cohen's alpha is 0.54 when using all module scores and 0.74 when dropping the score for Module 1.

Table 8. Cohen's alpha score when a module is dropped (data collected summer/fall 2021)

Module Dropped	*n*	alpha
none	167	0.54
1	**167**	**0.74**
2	167	0.41
3	167	0.37
4	167	0.28

For the efficacy study, we use IACT modules 2, 3, and 4 for analysis. This measure has an internal reliability (Cohen's alpha) of 0.64 for a large data set and 0.74 for a smaller data set more focused on neurodivergent students.

ACE Measures of Executive Function

We examine data for 205 students who completed UCSF's Adaptive Cognitive Assessment (ACE). 165 of these students had data for both ACE and IACT. ACE is a validated collection of cognitive tests designed for use across age groups, and has been used in studies of math achievement and the neuroscience of attention (Coulanges et al. 2021, Mishra et al., 2021). The specific tasks used appear in Table 9.

Table 9. ACE tasks included in study.

Task Name	Dimension Measured
BRT	Basic Response Time
Backwards Spatial Span	Working Memory
Flanker	Selective Attention
Task Switch	Task Switching
SAAT Impulsive	Response Inhibition

Scores for each task were calculated using the *aceR* package in R, developed by UCSF. While this package provides a wide assortment of possible metrics for scoring these tasks, we use those recommended by the developer: Rate Correct Score (defined as accuracy/mean response time) for Task Switch and Flanker, mean response time for BRT and SAAT Impulsive, and maximum object span for Backwards Spatial Span. We observed a normal distribution of scores for each task (Figure 7):

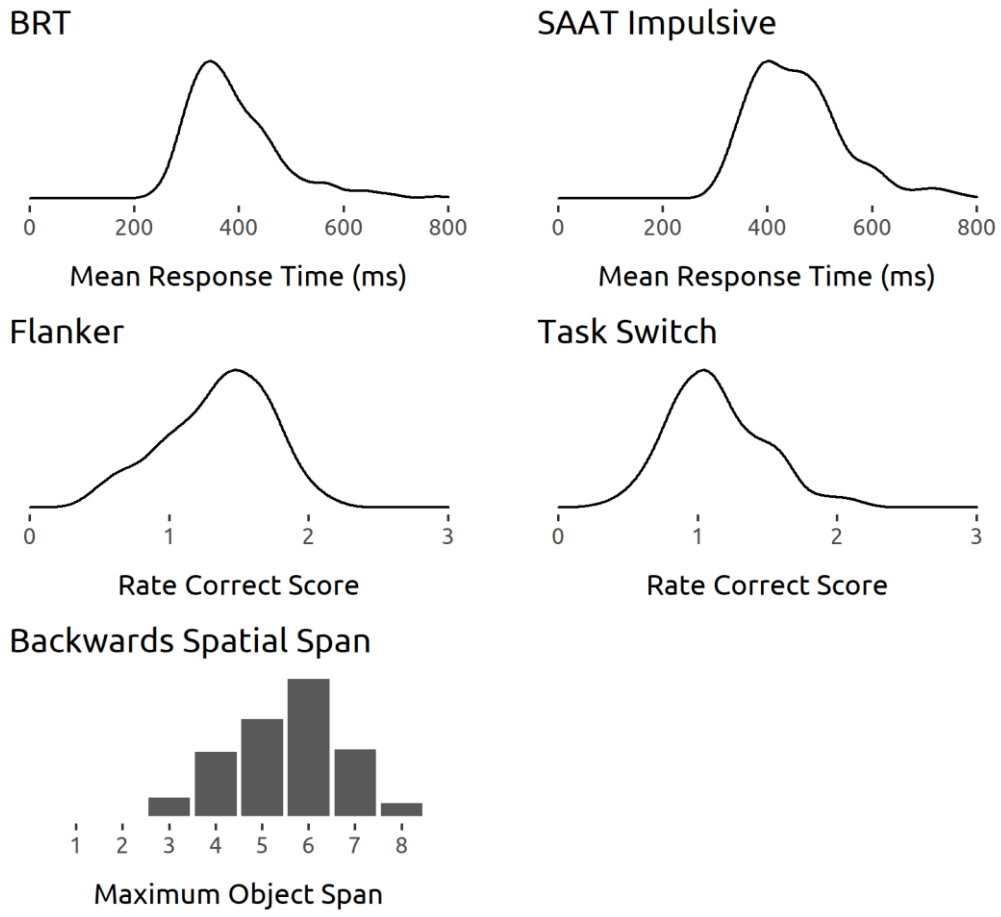

Figure 7. Distribution of scores on ACE measures, validation study (N = 205).

We found that scores on the various ACE tasks were weakly to moderately correlated with each other. Further, none had a correlation stronger than 0.10 with overall IACT score (mean of z-scores for Modules 2-4), indicating that **IACT is a suitable measure to use for populations with varying executive function proficiencies**. These results are shown in Table 10.

Table 10. Correlation coefficients for ACE and IACT measures, validation study (N=165).

	BRT	SAAT Impulsive	Flanker	Task Switcher	Backwards Spatial Span
Overall IACT	**-0.07**	**0.10**	**0.00**	**0.01**	**0.07**
BRT	1	0.52	-0.23	-0.31	-0.34
SAAT Impulsive	-	1	-0.30	-0.38	-0.20
Flanker	-	-	1	0.51	0.09
Task Switcher	-	-	-	1	0.11
Backwards Spatial Span	-	-	-	-	1

Linear regression indicates that the only ACE metric with a detectable relationship to the IACT score is SAAT Impulsive (p = 0.04), but as with the correlation coefficient (-0.10) the effect size is negligibly small (partial eta-squared=0.027).

Knology

Processes

Biosphere

Culture

Media

Wellness

Systems

Knology.org
fax: 347-288-0999

tel: (442) 222-8814
3630 Ocean Ranch Blvd.
Oceanside, CA 92056

tel: (347) 766-3399
40 Exchange Pl. Suite 1403
New York, NY 10005