INTERNAL USE

Development and validation of a game-based mathematics assessment using concrete representations of math
2021 IES/SBIR PROJECT REPORT *

Keith Devlin, Ph.D. (BrainQuake)
Howard Everson, Ph.D. (BrainQuake & City University of New York)
Bryan Matlen, Ph.D. (WestEd)
Randy Weiner, MA (BrainQuake)

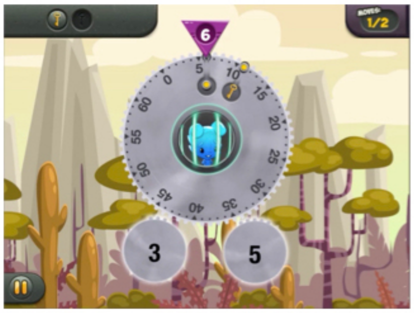
January 3, 2022

*A 3.5min video summary of the project can be viewed at https://vimeo.com/793683677

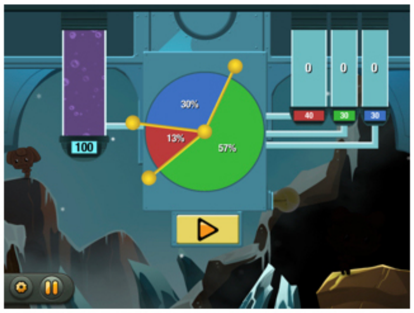
1. OVERVIEW

BrainQuake seeks to develop and validate an online assessment suite addressing three different aspects of the upper-elementary and middle-school mathematics curriculum (see screen images below):

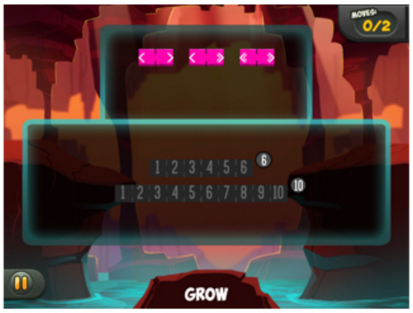
1) Proportional Reasoning & Fractions (based on the BrainQuake *Tanks* puzzle)
2) Arithmetic & Algebraic Thinking (based on the BrainQuake *Gears* puzzle)
3) Functions, Linear Growth, and Spatial Reasoning (based on the BrainQuake *Tiles* puzzle)



Integer arithmetic & algebra



Fractions & proportional reasoning



Functions and linear growth

We note that the main goal of all BrainQuake puzzles (and the more so for the entire suite taken together) is to develop the crucial, foundational 21st Century skills of number sense, multi-step reasoning, and creative problem solving, and hence *an assessment produced using them will provide valuable information on these critical math skills, which multiple-choice tests can miss*.

In this project, we built a prototype study assessment comprising ten items based on the *Tanks* puzzle (item 1 in the above list and central image) and its associated Digital Manipulative (see later), designed to explicitly support an assessment use case, and validated it against ten proportional reasoning items from the New York State Common Core Mathematics Test.

We envisage a follow-up project in which we replicate the validation protocol we applied to the *Tanks* puzzle in this project, to a suite comprising items based on all three puzzles listed above. Based on our observations from this study, we would double the number of puzzles of each of the three types to 20, and use a comparison standardized test of 30 questions, 10 each chosen to match each of the three BrainQuake puzzles in terms of mathematical content.

Note that a larger group of assessment items is required to formally validate each kind of item. It does not follow that use of the assessment tool would involve so many items.

Assuming success, by the end of such a follow-up project, we would have an online assessment of crucial and fundamental mathematical thinking skills with demonstrated correlation to standardized testing performance. As such, BrainQuake would be able to support the following four use cases:

1) **In-school Formative Assessment**. Assuming the correlation data we obtained holds or improves for all of our puzzles at the end of the second project, BrainQuake would be able to state that using our puzzle suite will provide predictive data regarding student performance on standardized testing items with respect to proportional and fractional reasoning; arithmetic and algebraic thinking; and functions, linear growth and spatial reasoning.

Teachers would be able to gather such assessment data on any schedule they like (as opposed to the NWEA MAP growth test, which schools typically only implement twice before an end-of-the-year growth measure) in order to receive insights into how their students are likely to perform on the standardized test topics BrainQuake's suite covers.

2) **At-home Formative Assessment**. Similar to the previous item, families, via our current consumer subscription SKU, would also be able to make similar use of their children's performance data on their own and/or in collaboration with their child's math teacher.
3) **In-school Practice Content**. In addition to supporting an assessment use case, the very same content can be used as practice content. This has traditionally been how BrainQuake content has been used in schools, as a warm-up exercise, the focus of individual or group problem-solving work and/or as part of a station rotation model. This flexibility extends our value proposition to schools
4) **At-home Practice Content**. Similar to the previous item 3, via our current existing consumer subscription SKU, families would continue to be able to use BrainQuake content as a fun and engaging way to develop math proficiency, without necessarily formally seeking assessment data on their children's performance.

2. PROJECT SUMMARY

Use of the BrainQuake learning items as an assessment required many changes and adjustments, based on a lengthy design period, but they did not result in any deterioration in the product's usability. We have robust web apps that did not crash a single time and that are deployable around the world; we have monitoring and logging in place that work flawlessly; and we are able to develop, very efficiently, new web apps that can flex in terms of number of assessment items and types. In addition, our cognitive interview data (see later) once again confirmed the ease-of-use and joy that students find in our work. Moreover, our prototype delivered actionable, predictive information with respect to standardized test items.

We developed and used teacher training materials to prep teachers, and we developed, for the first time, in-depth interface tutorials and practice items to support the assessment use case.

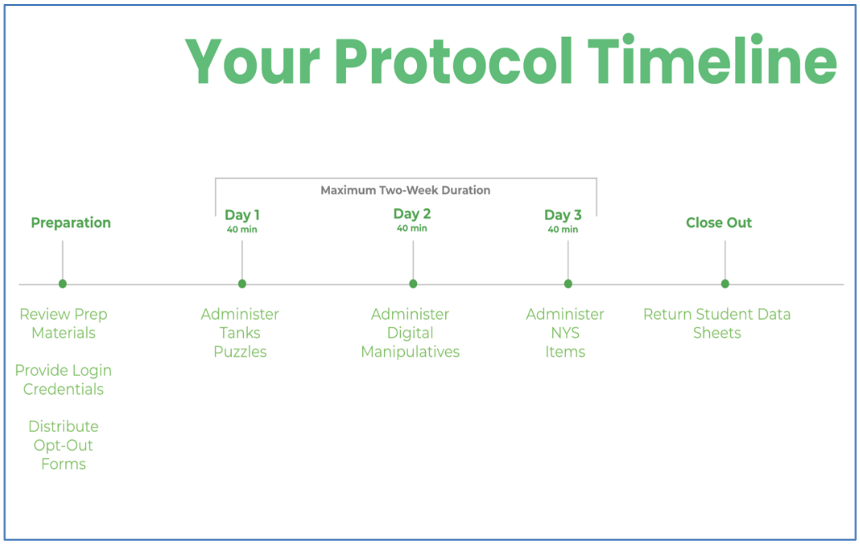
Building the prototype required nuanced and deeply meaningful iterations. Coming into this project, we had not designed our *Tanks* puzzle (or any of our puzzles) for a specific assessment use case. As such, we had to take a step back to determine what, for the purposes of a prototype, needed to change.

Building the prototype and conducting the validation study, required that we re-examine, and when necessary adjust, the many design decisions that had been made when we originally built the *BrainQuake* learning platform. This lengthy process was hugely strengthened by the participation of renowned assessment expert Dr. Howard Everson, who was brought in to be part of the project team. [The original development of the BrainQuake puzzles *as tools for learning* was based on the fundamental research carried out at Stanford by BrainQuake co-founder (and project P.I.) Dr. Keith Devlin on game-based math learning and the use of non-symbolic, direct representations of mathematics. Although BrainQuake's diverse team was able to build a successful learning product on that research base, *no assessment expert was involved*. Hence the recruitment of Dr. Everson.]

We list below some of the most significant changes we made to create the prototype and conduct the validation study. Since validation of the eventual product is crucial to its functioning as a *bona fide* assessment, it was essential that the prototype meet the demands of a formal validation process—something Dr. Everson has decades of experience in. Accordingly, virtually all design decisions were made explicitly with the assessment use case and validation study in mind.

Instructional materials and practice items

Our study protocol involved students taking three different assessments on three different days: puzzles, then digital manipulatives, then standardized test items. Prior to this study, BrainQuake explicitly did not seek to provide any interface instruction (beyond extremely lightweight, in-game tutorials) to users, since safe and engaging mathematical exploration is one of our core design principles. However, in order to decrease the likelihood of users failing to submit the answers they wanted to submit, we developed dedicated practice items for each item type (including a practice item for an additional Question 11 included with the New York State test—see later). We also created written and video content for teachers that explained not only how each item type functions, but also provided detailed, explicit instructions regarding how and what to teach students regarding each item type interface and functionality. The image shown right is an example of the guidance materials we created for the teachers.



Your Protocol Timeline

Maximum Two-Week Duration

Preparation

Day 1
40 min

Day 2
40 min

Day 3
40 min

Close Out

Review Prep Materials

Provide Login Credentials

Distribute Opt-Out Forms

Administer Tanks Puzzles

Administer Digital Manipulatives

Administer NYS Items

Return Student Data Sheets

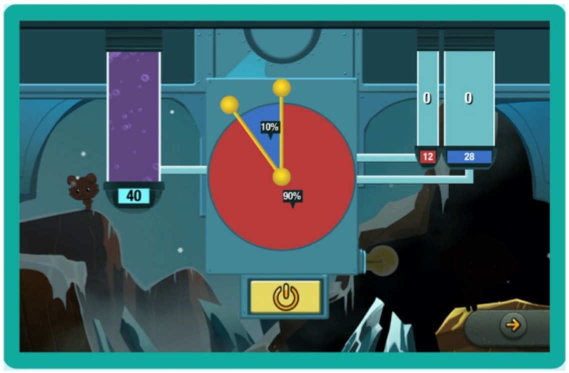
On each discrete day of testing, teachers provided students with access to the relevant practice problems (on the puzzle testing day, students practiced with the puzzle practice items to master the interface) prior to beginning the formal assessment. Practice items reflected the full range of functionality students might need to understand to demonstrate their math understanding, in order to minimize the possibility that student performance was impacted negatively by a failure to understand how to use each item type. This practice item “infrastructure” will serve us extremely well in a future project, when we extend our validation study to our remaining puzzles.

Selection of the assessment items

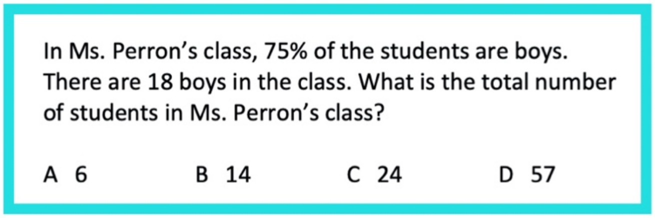
We began by selecting ten items on fractions and proportions from the New York State Common Core Mathematics Test for years 2013, 14, 15, for Grades 5, 6. We looked for problems that required conceptual understanding of fractions and proportions to solve (not just computational skills), since that is the primary focus of all BrainQuake learning products.

We then selected ten BrainQuake *Tanks* puzzles where the underlying mathematics was essentially the same, thereby permitting a valid comparison of the two assessments. The image below illustrates what we mean by “essentially the same” here.

BrainQuake Puzzle item #7



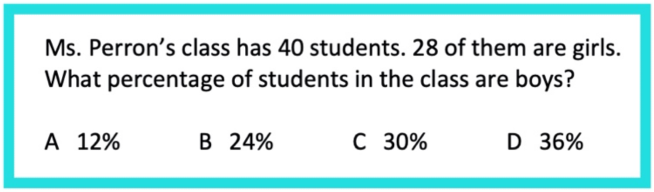
NYS Test Question #7



In Ms. Perron's class, 75% of the students are boys. There are 18 boys in the class. What is the total number of students in Ms. Perron's class?

A 6 B 14 C 24 D 57

Puzzle #7 formulated as a m.c. word problem



Ms. Perron's class has 40 students. 28 of them are girls. What percentage of students in the class are boys?

A 12% B 24% C 30% D 36%

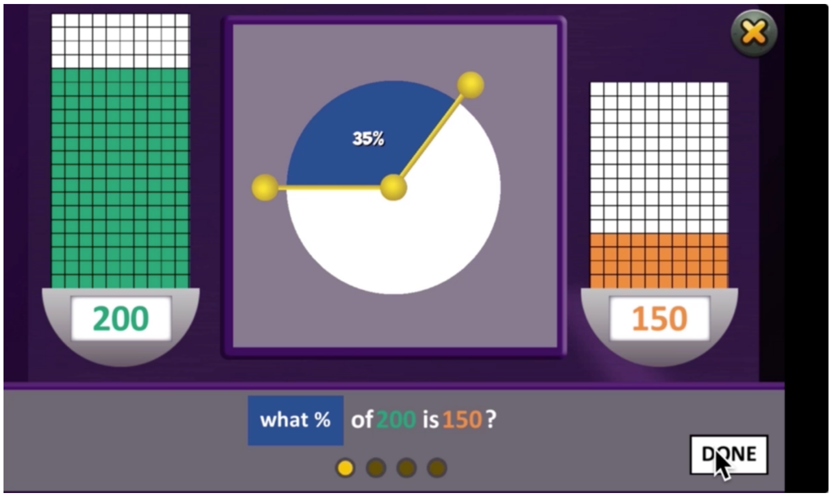
The top left image in this figure shows the 7th BrainQuake puzzle presented to students in the study, and next to it, the top right image shows the 7th New York State question they saw. They look very different. But when we translate the BrainQuake puzzle into a multiple-choice, symbolic-math question, tailored to the same word-problem of the New York State question, as shown in the image on the bottom right of the figure, we see that the two symbolic questions are virtually identical They are clearly mathematically equivalent.

As this example highlights, although there are indeed differences between the two tasks—that's the whole point of the BrainQuake approach of breaking the symbol barrier—the underlying mathematics can (and does) remain essentially the same.

Inclusion of the *Tanks* Digital Manipulatives

The study also presented students with ten DMs associated with the *Tanks* puzzle. (See image below right.) We did so with two research goals in mind:

- To provide information that might help us understand any similarities or differences between student performances on the BrainQuake puzzles and the NYS items. (The DMs bridge the gap between the two, by combining dynamic, symbolic math expressions with the puzzles.)
- To study the degree to which the DMs themselves can function as assessments, and how well they would be validated against the NYS items. (Validation would mean we could include DMs in our future assessment product, if doing so resulted in a demonstrably better assessment.



The study raised a number of questions concerning student performances on the DMs and an 11th question we added to the New York State test (see later) that will require consideration and study in the follow-up project. We anticipate that answering those questions will require increasing the number of items we administer (from 10 to 20 for the puzzles, perhaps not such a large increase for the DMs and standardizes test items), since game-based learning depends upon achieving a student flow, which is achieved by starting with a slow, gentle ramp. (We would expect the actual assessment to be computed from the final ten puzzles, with some of the first ten preparing the groundwork for the second ten.)

Inclusion of an experimental test item to probe student conceptual understanding of symbolic math

The New York State test module we presented to students also included a Question #11, which comprised ten YES/NO questions, expressed in symbolic format, designed to measure students' *understanding* of fractions (e.g. the different role of numerator and denominator), which students were asked to complete as quickly as they could. None of the individual items required any computation to provide an answer. The goal was to probe the same question of the students' ability to see beyond math's symbols to the concepts they represent as the DMs do, but whereas the DMs do this with the conceptual game mechanics as a starting point and inject symbols, the Question 11 took a traditional-looking multiple-choice (actually, binary choice) test as the starting point but focused exclusively on the concepts the symbols represent. As with the DMs, the decision to include this item was to provide another mechanism whereby we might be able to analyze the results from the primary validation study, by identifying any differences in student performance on the BrainQuake items and the New York State items.

Size of the tests

In our study, we ultimately implemented 10 puzzles, 10 digital manipulatives and 10 standardized test items (plus the additional, experimental standardized test item). We engaged in much iterative discussion and consultation regarding how many items we could reliably expect students to finish in a single 40-minute class period (items of each of the three type—puzzles, digital manipulatives and standardized test items—were administered on separate days). In an ideal world, we would have liked to have administered more items of each type, but we felt the risk to completion was too high, and we did not want to find ourselves in a position with a wide range of varying levels of item completion in our data set. On the other end of the spectrum, we considered administering as few as seven items of each type, feeling extremely confident that there would then be virtually no risk to completion on each of the three days.

We ultimately settled on 10. Though we did not find this decision to be terribly satisfying in terms of the science, we did feel it was reasonable and, even more important, necessary to keep the size of the tests limited due to the impact COVID continues to have in schools. Most transparently, we are unendingly grateful that any teacher anywhere in these last months made time for us. We added an entirely new element into a chaotic and stressful school environment and the teachers really had no reason to take on anything new in the face of the operational and emotional complexity schools are facing.

In a future project, assuming more stable conditions in schools, we would seek to implement our more ideal, 20-item validation scenario. Note that the need for so many items of each type is required to *validate* each type. We do not see a need for so many items in order to assess student performance in the final product.

Puzzle and digital manipulative design and functionality

Prior to schools opening in the fall, we engaged in a significant amount of internal discussion, debate and prototyping regarding what we would ultimately name as a critical feature of our assessment items: *replayability*.

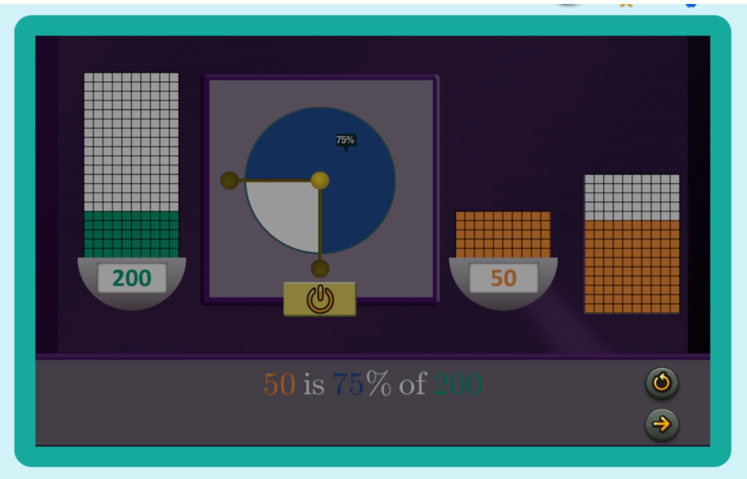
Standardized tests generally do not include any form of feedback. Once a student selects and submits an answer, the next item is presented. Little, if any learning, is possible via the experience of taking the assessment. Though we did take a few moments to examine whether or not to implement replayability for the standardized test items in our study, we ultimately decided not to do so in order to maintain the utmost integrity of how students experience these items in real world testing situations.

Our approach to replayability with regard to our puzzles and digital manipulatives swung back and forth multiple times during the months of May through September. On any given day, different members of our team could sway our collective opinion in favor of no replayability ("we should minimize variability between the standardized test items and the BrainQuake items") or in favor of it ("replayability, and the learning opportunities that we create through it, are an essential component of BrainQuake's design principles and learning philosophy, and a compelling new value proposition for assessment" and of game-based learning in general).

Ultimately, we chose to support limited replayability, allowing up to a total of three attempts on a given puzzle or a given digital manipulative. This limited replayability is unique to our assessment product, as all of our original designs support unlimited replayability. By supporting replayability, we allowed ourselves to collect potentially disruptive data: if we found that students demonstrated comparable proficiency to their performance on the standardized test items based upon scores derived from a second or third replay, what should we then ultimately say about that student's proficiency? Does eventual proficiency disqualify or discount failure on the first attempt? If not, then what should we conclude about students who do not demonstrate proficiency on standardized tests that only support one attempt? (We note that the multiple-choice format of most standardized test questions cannot tolerate multiple attempts and remain a valid assessment. In contrast, having gotten a BrainQuake item wrong, in trying a second time the student is right back at the starting point, since the items are open-ended; in most cases the solution space has at least hundreds of pathways and often millions.) These are issues we have not yet had an opportunity to analyze, but we intend to do just that later this month (January).

Supporting limited replayability also required us to design and iterate on new feedback mechanisms, as well as re-consider our existing forms of feedback. As always, we wanted to limit our use of written language to the greatest degree possible, as overcoming language barriers is another essential BrainQuake design principle. For our puzzle items, we reduced our feedback on unsuccessful tries to just 11 words. Regarding the Digital Manipulatives, we used just two words.

We iterated further, though, on the Digital Manipulatives feedback. Prior to this project, we did not provide comprehensive feedback to users on an unsuccessful solution. Here, though, for the assessment use case, given that we chose to support replayability, we wanted to make sure that students were actually able to self-assess what might have gone wrong with the intended solution. As such, we designed and implemented the ability to fully present on screen the complete representation of an over- or under-filled output tank. This proved to be more challenging than we initially thought, as we had to develop a solution that temporarily shrunk all on-screen assets to accommodate edge cases that required more screen real estate than our original design provided. The figure below right illustrates our approach.



50 is 75% of 200

The screenshot also presents another new aspect of digital manipulative feedback. The screen appears dimmed while the buttons are brighter. Unlike in our previous, non-assessment design, here we are “freezing” the screen after an answer submission. This allowed students to control how much time they wanted to review the outcome of their previous attempt. When ready, students could choose to either skip the item or replay it by clicking the circular arrow button. Supporting this level of agency is yet another core BrainQuake design principle that we were able to instantiate in a new and different way as part of our assessment design.

Puzzle success feedback

We discovered that, for the purposes of our assessment use case, we needed to remove the star scoring system that we use in our consumer SKU. We did experiment with leaving the star system in place, but ultimately we decided that it was an unnecessary variable that had no similar counterpart in the digital manipulatives’ or standardized test items’ success. In that we wanted to minimize variability across items types as much as possible, we removed the star system and updated our code-base to reflect not only that change, but also to strip out associated feedback and navigation elements that accompany the star system in the consumer SKU, such as displaying stars earned on the map and presenting a summary puzzle performance pop-up following the completion of a given puzzle.

Skip functionality symmetry

Following our rationale regarding maintaining as much consistency across item types as possible, we also implemented the same skip option for each item. Students could skip any puzzle they wanted, but unlike in our consumer SKU, where we do not “challenge” the student’s decision, here in the assessment use case, we wanted to encourage students to provide as many answers as possible in the name of creating the most robust data set possible. As such, all item types had the same skip functionality. Reading about such an iteration here on the page may make the iteration seem small and minor, but these kinds of details reveal themselves as exceedingly important to the overall integrity of the study, and they only reveal themselves as a result of iteration. (Much of the power of game-based learning comes from dealing with learning—and now assessment—issues by the use of good system design.) We will revisit issues such as this in a future project.

Dedicated item-type maps and the three-day testing protocol

Our initial conception of how we would deploy the assessment items imagined a single BrainQuake map with all item types, with testing occurring on just one day. However, as we engaged in discussion regarding how many items of each type we wanted to include, what schools’ assessment administration constraints would look like, and how we could make the administration as simple as possible for the participating teachers, we determined that each item type should live on a single map ***and*** that each item type should have its own testing day. As such, we leveraged our modular architecture to develop and rapidly deploy three discrete maps, one for each item type and its associated testing day. Teachers were encouraged to administer the tests in succession over three days (but no longer than within a two-week period), following this order: puzzles, digital manipulatives, and finally standardized test items on day three. This is another instance where the requirements in terms of student and teacher time and number of items to be administered for a validation study greatly exceed what would be required to assess a student's performance in a typical assessment use-case.

Monitoring, logging, and scoring

As our assessment protocol evolved, so too did our need to iterate on how we monitored, logged and (internally) scored student interaction with each item type. As is always the case with monitoring and logging, the challenge is not how to do it, but how much of it to do. Despite our eagerness to capture all manner of data, we ended up capturing a very small and focused subset, relevant to the study, of all available data. This included number of attempts per item, elapsed time per attempt and across all attempts, answers submitted per attempt, items skipped, skips selected in accordance with attempt number, correct answer recognition, incorrect answer recognition and incomplete answer submission.

3. PILOT STUDY TO TEST THE USABILITY AND FEASIBILITY OF THE PROTOTYPE

Sample of Students The sample consisted of 238 5th and 6th graders recruited from middle schools in California, New York, and Michigan. In addition to working on the BrainQuake *Tanks* puzzles and digital manipulative tasks (the DMs), students were also administered a 10-item multiple-choice test designed to measure state-mandated math standards for proportional reasoning identified for 5th and 6th graders. The 10 items we selected for use were drawn from a pool of released items made available by the New York Education Department. An 11th item was included (for use as a researchers' analytic tool), comprising 10 YES/NO questions about fractions that students were asked to answer as quickly as possible.

All the assessment tasks were administered online and scored using a simple number-correct scoring procedure. In addition, teachers provided the researchers with indicators of the students' proficiency in mathematics and English language using a 1-5 rating scale with 1 indicating below minimum proficiency, 2 = minimally proficient, 3 = basic proficiency, 4 = advanced proficiency, and 5 = very advanced proficiency levels. All student level-data, which includes the puzzle-based performance data, data from the innovative DMs, students' scores on the proxy math test, and school-based proficiency classifications, were collected from the students during the Fall of the 2021 academic year.

Sampling Issues Unfortunately, two of the targeted school districts backed out due to teachers being overwhelmed by the pandemic and time constraints. The research team then did extended outreach to additional schools and districts and were able to secure participation from several new sites—including schools in Michigan and California. The process was to contact a primary contact in the school district leadership (or charter management organization) to get initial buy-in and for them to assist in recruiting teachers. An MOU was signed by the district/charter leader and teachers volunteered to participate by signing a consent form. Each teacher was offered a $250 stipend for their effort.

Each participating teacher attended a 50-minute orientation on the purpose of the study and training session on how to implement the study. That session followed a script that explained the background of BrainQuake and the online platform, how to assign passwords to students, the theory behind the design of the puzzles and digital manipulatives, and a detailed walk-through of each assessment (how students accessed it and how it worked). Teachers also learned how to complete and return the student demographic data sheet, and were given a timeline for completion of the study and the submission of student-level data.

The teachers distributed all the required consent forms to parents and students via email and a "SurveyMonkey" link was created so the consent forms could be submitted to the research coordinators at each site. Participating students' names and other personally identifiable information were deleted by the teachers prior to submission to the researchers for analysis.

Teachers had all the students in their classroom(s) complete the study. That varied from a single contained classroom for the 5th grade at small schools, to multiple sections of 6th grade math (up to four classes per teacher). Each session followed a sequence of starting with an explanation of the platform (how to log in), a practice assessment module to learn how it works, then the actual assessment. Each class started with the BrainQuake *Tanks* puzzles on the first day, then the BrainQuake DMs the second day, and finally the standardized math-assessment items. Most teachers completed the study in three consecutive days in one week, though a few had to carry over into the following week. The study settings were the teachers' classrooms, with no contact from the research team. All students had their own computer (usually a school Chromebook) and were provided their own log in credentials.

The Sample The students and teachers who participated in the study were recruited from three school districts across the U.S.—the Rapid River School District in Michigan (N=14), the International School of Monterey in California (N=23) and the Wantagh School District in Long Island, New York (N=201). Although smaller than anticipated, the sample of students (N=238) was, broadly speaking, representative of the middle school students—fifth and sixth graders who ranged in age from 10 to 12 years old. Tables 1 and 1A provide summarizes of the key demographics and learner characteristics of the students in our sample.

We note that the project team is continuing work on the project unfunded to try to bring in an additional 80 to 100 students, largely of a Latino demographic, to add to the application data for our next project. (The teacher and school had committed, but ran into technology issues that they were unable to resolve in time.)

Table 1. Sample Descriptive Statistics: Age, Gender, Race, SES Proxy, English Language Learners (N = 238).

```
      Age |      Freq.     Percent        Cum.
       10 |         26       10.92       10.92
       11 |        165       69.33       80.25
       12 |         47       19.75      100.00
   ------+----------------------------------
   Gender |
    Female |        126       52.94       52.94
     Male |        112       47.06      100.00
   -------+----------------------------------
    Grade |
        5 |         37       15.55       15.55
        6 |        201       84.45      100.00
      -------+-----------------------------------
       Race |
       White |        199       83.61       83.61
       Black |          2        0.84       84.45
      Latinx |         12        5.04       89.50
 Asian Amer.|         13        5.46       94.96
Native Amer.|          4        1.68       96.64
       Other |          8        3.36      100.00
```

Table 1A. Sample Descriptive Statistics: SES Proxy, English Language Learners Status, and English Language Arts and Math Proficiency Levels.

```
   Students |
  SES Proxy |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        211       88.66       88.66
          1 |         27       11.34      100.00
------------+-----------------------------------


   Students |
 ELL Status |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        215       90.34       90.34
          1 |         23        9.66      100.00
------------+-----------------------------------


      Students ELA |
 Proficiency Level |      Freq.     Percent        Cum.
-------------------+-----------------------------------
               Low |         14        5.88        5.88
           Minimal |         37       15.55       21.43
        Proficient |        102       42.86       64.29
Clearly Proficient |         59       24.79       89.08
          Advanced |         26       10.92      100.00
-------------------+-----------------------------------


     Students Math |
 Proficiency Level |      Freq.     Percent        Cum.
-------------------+-----------------------------------
```

```
               Low |         14        5.88        5.88
           Minimal |         54       22.69       28.57
        Proficient |         75       31.51       60.08
Clearly Proficient |         59       24.79       84.87
          Advanced |         36       15.13      100.00
-------------------+-----------------------------------
Proficiency                Freq.     Percent        Cum.
-------------------+-----------------------------------
```

```
          2 |              8        3.36        3.36
          3 |             10        4.20        7.56
          4 |             28       11.76       19.33
          5 |             24       10.08       29.41
          6 |             66       27.73       57.14
          7 |             28       11.76       68.91
          8 |             35       14.71       83.61
          9 |             17        7.14       90.76
         10 |             22        9.24      100.00
```

The data in Table 1A, which summarize the teachers' estimates of their students' proficiency and indicate that, according to the teachers, about 34% of the students are less than proficient in Math, and roughly 28% are less than proficient in English. When we look at the distribution of scores after computing the joint proficiency classification, we see that roughly one-third of the students are less than proficient in both Math and English, and another one-third are classified as having only basic levels of proficiency in both domains.

The two key research objectives focused on here are: (1) to fit a series of statistical models to estimate the puzzle-based tasks predictive value when assessed against students' math scores after controlling for students' age and overall proficiency levels; and (2) to provide empirically based evidence of the validity of the puzzle-based tasks for estimating students' proportional reasoning ability in math for this sample of 5th and 6th grade students. The results of our analyses are presented next.

Results

Table 2. Summary of Students' Performance on All Four Math Assessment Tasks

Variable |	N	Mean	SD
Math Items |	221	4.20	1.79
BQ DM Tasks |	145	1.95	1.40
BQ Puzzles |	227	5.24	1.85
BQ Tasks |	139	7.52	2.59

Note: The math tasks, the DMs and the Puzzles were scored on a 0-10 scale;

the BQ Tasks (DMs + Puzzles) were scored on a 0-20 scale.

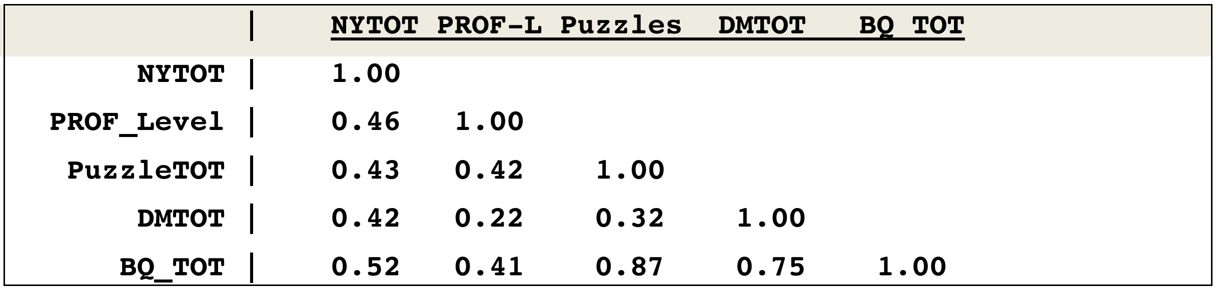
As we see in Table 2 the students in our sample averaged 4.2 items correct on the New York State multiple-choice items and averaged 5.2 BrainQuake puzzles solved correctly.

The students did not perform well on the BrainQuake digital manipulatives (DMs), correctly solving on average less than 2 out of 10 tasks. When we summed the BrainQuake tasks—puzzles and DMs—the scores indicate that, on average, the students correctly solved about 7 or 8 tasks out 20.

The students also performed poorly on the 11th question included with the New York State items, but its purpose was to provide an additional source of information for the researchers to inform future work, and is not included in the analysis reported here. (Though the students' poor performance on that task does provide a clue as to their poor performance on the DMs, suggesting that the juxtaposition of a game-puzzle along with a symbolic math question may be what causes problems for learners. The DMs were designed purely as a learning tool that teachers could leverage to help students transfer skills acquired during the game to performance in traditional classroom math. Their potential for use as part of an assessment remains an open question of great potential significance, that we intend to pursue in the next project.)

To investigate further the relationships among and between the four math measures in Table 2, we analyzed the scores by estimating the zero-order correlations among them, as well as by conducting a series of multivariate regression analyses. We turn to these analyses next.

Table 3. Correlations of Students' Math Scores, Proficiency Levels, Puzzle Scores, DM Scores on the Combined BrainQuake Tasks Scores.



```
             |    NYTOT  PROF-L  Puzzles  DMTOT    BQ TOT
       NYTOT |    1.00
  PROF_Level |    0.46   1.00
   PuzzleTOT |    0.43   0.42    1.00
       DMTOT |    0.42   0.22    0.32    1.00
      BQ_TOT |    0.52   0.41    0.87    0.75    1.00
```

Table 3 presents the correlations among and between five key measures on interest—the standardized math items (NYTOT), the students' overall proficiency levels which combines the teachers' proficiency estimates in both Math and English Language Arts (PROF_Level), the puzzle scores (PuzzleTOT), the DM (DMTOT)scores, and the total of both the puzzles and the DMs (BQ_TOT). We included the proficiency level scores because those indices, unsurprisingly, were moderately correlated with performance on the math items (r = .46) as well as on the puzzles (r = .43), the DMs (r = .42), and the BQ total score (r = .52).

We looked more closely at this pattern of correlations by fitting a series of nested regression analyses to isolate the evidence of the predictive value of the BrainQuake tasks, both the puzzles and the DMs separately and together, when it comes to estimating students' performance on the standardized math items (our criterion variable for purposes of predictive validity modeling).

Regression Analyses

This section presents the results of our analysis of the seven competing regression models we tested—all of which attempt to predict performance on the criterion measure, i.e., the 10-item multiple-choice math test administered to the sample of students. Table 4. describes the components of each of the models beginning with the baseline model that simply predicts students' performance on the math test as a function of their overall proficiency levels. Models 2-4 simply substitute the three BrainQuake scores for the students' proficiency levels, and Models 5-7 include both the students' proficiency levels and each of the three BrainQuake scores in an effort to isolate statistically the predictive value of those scores after controlling for the students' proficiency levels.

Table 4. Description of Seven Competing Regression Models.

Regression Models	Prediction Variables
Model #1	Proficiency Level
Model #2	BQ Puzzle Score
Model #3	BQ DM Score
Model #4	BQ Total Score
Model #5	Prof. Level + BQ Puzzle Score
Model #6	Prof. Level + BQ DM Score
Model #7	Prof. Level + BQ Total Score

The seven models listed in Table 4 were analyzed sequentially and the model fit indices resulting from those analyses are summarized in Table 5.

Table 5. Summary of the Model Fit Indices for Seven Competing Regression Models Predicting Students' Performance on the Ten Item Standardized Math Test.

	M1	M2	M3	M4	M5	M6	M7
Mult.*R*	**.47**	**.43**	**.42**	**.53**	**.53**	**.55**	**.58**
R^2	**.22**	**.19**	**.18**	**.27**	**.28**	**.30**	**.33**
Adj.R^2	**.21**	**.18**	**.17**	**.26**	**.27**	**.29**	**.32**
RMSE	**1.59**	**1.64**	**1.67**	**1.59**	**1.55**	**1.54**	**1.54**
N Obs.	**221**	**211**	**138**	**132**	**211**	**138**	**132**

The indices in Table 5 include the multiple *R* (a measure of the correlation), the R^2 and the adjusted R^2, indices of how much of the variation in the criterion variable (i.e., the math scores) is explained by the regression model, the Root Mean Square Error (RMSE)—an estimate of the prediction error in the models, and the number of observations included in each of the statistical analyses. Generally, when comparing the competing prediction models, we are looking for increases in the R^2 indices and decreases in the RMSE indices.

Conclusions

The summary statistics presented in Table 5 are very encouraging with respect to the evidence in support of the predictive validity of the BQ Puzzles and DMs, once we control for the students' initial levels of proficiency in Math and English. This can be seen most clearly when contrasting Models 1-3 (the models with only a single predictor variable) with Models 4-7—the models that capture increases in predictive validity associated with the BrainQuake measures after statistically controlling for the students' initial levels of proficiency in Math and English. In particular, Models 5-7 indicate that, even after controlling for the students' overall proficiency levels reported by their teachers, the two BrainQuake measures (the puzzles and the DMs) add significantly to the prediction of the standardized math scores— R^2s of .28 and .30, respectively. The validity evidence in support of the BQ measures grows stronger once we include the overall BQ total score as we see in Model #7—the overall best fitting, most explanatory regression model.

We set out in this study to gather data from a large, representative sample of 5th and 6th graders. The COVID-19 pandemic, unfortunately, interfered with our plans. We did manage, however, to sample 238 students, most of whom (85%) were 6th graders and most of whom were not viewed as strong academically by their teachers. (We noted above that we may have results for an additional 80 to 100, mostly Latino students in time for inclusion in a follow-up project.)

The relatively low levels of proficiency in our sample served to restrict the range of performances on the four assessments intended to measure proportional reasoning. These students, for example, appeared to find the BrainQuake digital manipulative tasks very challenging, completing successfully only two of the ten tasks on average. Nevertheless, the results of the regression analyses summarized in this report are very encouraging with respect to demonstrating strong, promising evidence in support of the predictive validity of the BrainQuake measures.

Clearly, there is evidence of a meaningful association between what is measured by standardized, multiple-choice type math assessments and the BrainQuake puzzles and digital manipulative tasks.

Cognitive Interview Testing

Usability Procedure and Participants Seven 6th grade students and three 5th grade students participated in BrainQuake cognitive interviews during December 2021. The interviews were conducted by WestEd. Nearly all students attend school in the San Francisco Bay Area, aside from two students residing in the Portland, Oregon, and New York City metropolitan areas.

All usability sessions were conducted virtually using Zoom. Researchers conducted the usability session procedures as follows:

- Researchers reviewed the study details with students and confirmed interest in participation.
- Students performed a think-aloud exploration of the BrainQuake application, completing Tanks puzzles 1 through 10 (See Figure 1). Of the ten puzzles:
    - Four puzzles involved percentages.
    - One puzzle involved fractions/segments with labeled segments (i.e., 1/7, 2/7, etc.)
    - Four puzzles involved fractions/segments with unlabeled segments.
    - One puzzle involved decimals.
- Students were asked to think aloud session while completing the tasks to demonstrate students' approach to solving each puzzle.
- Researchers conducted a post-interview to further elicit student opinions and to determine possible areas for improvement.
    - The post-interview included questions on what students liked and disliked about the game, which parts of the game were easy or hard to understand, which types

of puzzles students preferred, whether the student would recommend the game to a friend, and suggestions for possible changes.

  - Researchers also asked students to describe the meaning of and relationship between the input tank, central device, and output tank.

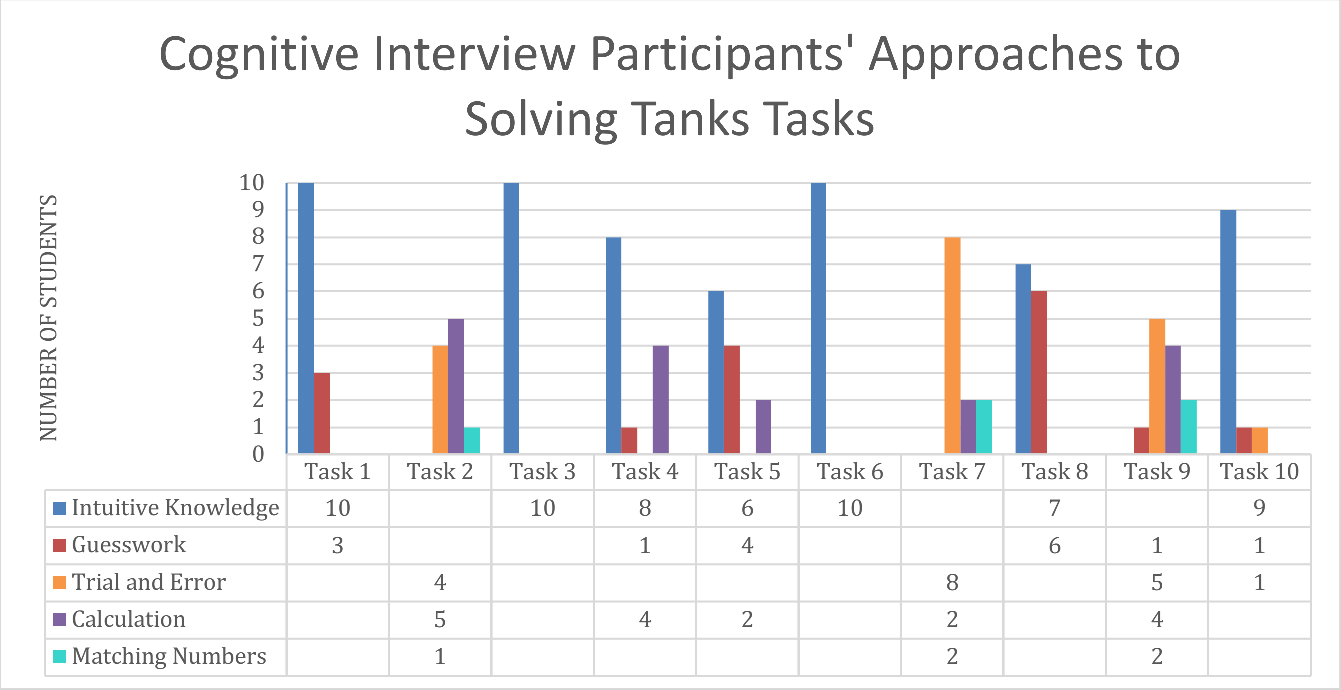
Fig. 1. Description of each Tanks puzzle task

	Input (left) tank amount	***Output (right) tank amounts***	***Question type***	***Amount of times input is needed***
Task 1	100	50, 50	Percentage	Once
Task 2	200	120, 80	Percentage	Once
Task 3	100	40, 60	Percentage	Once
Task 4	200	50, 150	Segmented, no labels (quarters)	Once
Task 5	200	100, 300	Segmented, no labels (quarters)	Twice
Task 6	70	30, 40	Segmented, labels (sevenths)	Once
Task 7	80	48, 32	Percentage	Once
Task 8	45	45, 45	Segmented, no labels (thirds)	Twice
Task 9	50	15, 35	Decimal	Once
Task 10	60	120, 120	Segmented, no labels (quarters)	Four times

Findings

Mathematical Reasoning

- Students had several different approaches when presented with solving the Tanks puzzles. Figure 2 breaks down the most common methods students used to solve each task. Each approach is described in further detail below.



	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Task 7	Task 8	Task 9	Task 10
🟦 Intuitive Knowledge	10		10	8	6	10		7		9
🟥 Guesswork	3			1	4			6	1	1
🟧 Trial and Error		4					8		5	1
🟪 Calculation		5		4	2		2		4	
🟩 Matching Numbers		1					2		2	

Fig. 2. Student solving methods.

*Note: Students often used more than one approach when solving a puzzle.

- **Easily intuitive puzzles**
  - All students were able to figure out how to move the arms of the central device and submit an answer via the power button fairly quickly, and without prompting. However, the majority of students verbally and/or behaviorally indicated confusion when encountering the first puzzle. One student almost pressed the skip button instead of the power button because they believed that the power button would end the game.
  - Despite this initial confusion, students were able to glance at many of the puzzles and solve them quickly. While some students said that they used quick mental math to solve the puzzle, researchers determined that most students relied on their intuitive knowledge as a method to solve such puzzles.
  - All students used this method to solve Tasks 1, 3, and 6. These three puzzles could be solved by matching the values of the tank on the right to the tank in the center. Most students relying on their intuitive knowledge to solve puzzles were able to solve the puzzle in one attempt.
    - *"It took me a second to figure it out, but then it was pretty easy to figure out what you were supposed to do." – 6th grade student solving Task 1*
  - Students most frequently struggled when encountering novel puzzle features (i.e., puzzles with unlabeled segments, puzzles with decimals, puzzles in which the power button must be pressed multiple times, etc.). However, students were generally successful in applying their experience from previous problems to subsequent problems, and showed improvement when encountering these features for the second time.
  - An example of this is Tasks 8 and 10, which involved a center tank that was segmented into thirds that required students to press the "submit" button more than once. Although around

half of students made guesses when approaching Task 8, 9 of 10 students were able to intuitively solve Task 10. This demonstrates that, after figuring out how to solve one type of puzzle, students were able to transfer that knowledge when presented with a similar type of puzzle later on.

    - *"So since I understand how to do [Task 8] now, this one is easier for me." – 6th grade student solving Task 10*

- **Guesswork**
  - Students made guesses when approaching some puzzles, especially Tasks 1, 5, and 8. A few students guessed on Task 1 because they were becoming familiar with the game and didn't understand how to navigate the puzzles without instructions. Tasks 5 and 8 required students to press the input button twice, which was not intuitive to many students.
    - *"I have no idea how I got that but I got that. I just clicked around and it worked." – 5th grade student solving Task 5*
- **Matching Numbers**
  - In Task 1, the tank on the left showed "100," so students could match up the numbers of the tanks on the right to the numbers on the pie chart in the center to solve the puzzle. A few students attempted to use this method when solving other problems, thinking that it would work. Several students expressed confusion when using this method for Task 7, in which the total amount of the left tank was out of 80.
- **Trial and Error**
  - Some students made an initial educated guess to solve the puzzle. After seeing the amount that one of the tanks on the right overflowed, students adjusted their initial guess and submitted the answer again. While students could eventually solve the puzzle correctly using this method, it almost always took them over 3 attempts to do so.
    - *"I just want to see how much this [tank overflows], so that I can adjust it." –6th grade student solving Task 9*
    - *"Since [when I submitted] the one before I saw that the red was a little less, I just put a little more to the red and it turned out to be [the] right [answer]." –6th grade student solving Task 7*
- **Calculation**
  - Students were told at the beginning of the session that they could use a pencil and paper to solve the puzzles if they wished. While no students opted to use this method, several students explained the calculations they made in their head when approaching a puzzle. Students more often used this method for Tasks 2, 4 and 9, where the total amount of the tank on the right was either 200 or 50, numbers that students said were easy to equate to 100 in their heads. This allowed them to easily convert the numbers on the right tank into the answer that they input into the center tank.
  - A few students used a "mental manipulation" approach for more difficult problems, such as Task 7, where the left tank's total amount was 80 and not easily divisible into 100. Students broke down the other numbers in the problem to make it easier to solve.
    - *"Since 15+35 is 50, I was just thinking of doing 15 times 2 and 35 times 2, because 50 is only half of 100." – 6th grade student solving Task 9*

    - *“So 48+32.. That's 80… Oh yeah. Ok, so if you… since 8 times 4 is 32, and 8 times 6 is 48, I think that has something to do with it. I think it would be [40% and 60%]? I'll try that. Yeah, I thought it would work.” – 6th grade student solving Task 7*

Conceptual Understanding of Task

- Many students were able to create rich analogies of the relationship between the input tank, central device, and output tank, demonstrating a conceptual understanding of proportional reasoning.
  - Though no students used the term “proportion” to describe the meaning of and/or relationship between the input tank, central device, and output tank, many were able to create rich and inventive analogies in its place. These students commonly viewed the input tank as containing some sort of liquid (i.e., grape juice, fuel, etc.) that was divided by the central device and then sorted into the output tanks.
    - *"When I think about it, I imagine it as one whole glass of grape juice. And then if you have a friend over, you're going to have to share that grade juice, so you would split it in half. The [Central machine represents] an amazing machine that will split these two juices, and it's like an automatic card shuffler, where it spits out two piles." –6th grade student*
    - *"[The input tank] is like the whole. [The central device] is like the tool you have to measure out, like if you cook or bake, [it’s] like the measuring cups. [The target tanks] tell you how much you need and what the quantity to quantity is. [You have to] figure out what you need in the central to equal the two smaller tanks to equal the bigger tank." –6th grade student*
    - *"[The three features] represent the fuel that is needed, the percentages you need to get the fuel, and the amount of fuel needed to light up the light. You take the amount of fuel you need and move levers around so you can get the right amount of fuel to even out the tanks, for the light to get on. [It’s like] traveling fuel from one end to another." –6th grade student*
- Other students had more disjointed explanations to describe the relationship between the input tank, central device, and output tank, but still demonstrated understanding of an underlying connection.
  - Of the students that did not provide obvious analogical reasoning, several still demonstrated knowledge of an underlying connection between the input tank, central device, and output tank. Instead of describing the input tank contents as a liquid, these students sometimes depicted the input tank as an indicator of what action one was supposed to take (i.e., how many times you would need to press the power button).
    - *"I feel like the purple [input tank] represents how many times you have to press it. [The central device] probably represents that it connects...like the purple [input tank] tells you how many times you have to press and [the central device] will help you get the amount. [It's all connected together because] it tells you how many times you have to click and [the central device] will tell you how much you have to put to get the answer." –6th grade student*
    - *“[The input tank] represents the amount you need to put in these two tanks. [The central device represents the] amount that it channels it. If you put an incorrect amount it would overflow.” –5th grade student*
- A few students had difficulty explaining how the parts of the puzzle were connected may have struggled to envision the three features as a cohesive unit.

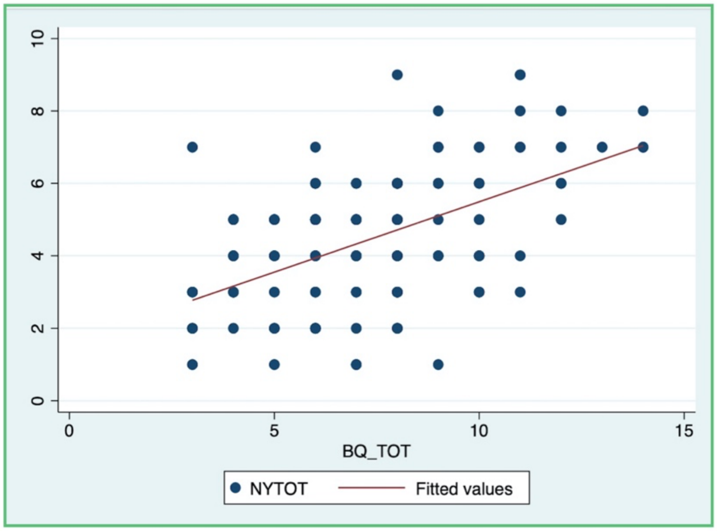
    - These students may have struggled to envision the three features as a connected unit, and likely had limited experience with the concept of proportional reasoning.
        - *"The [input tank] makes it to 100, and you try to make that on the wheel. I was confused why it didn’t go on there with the 61. And the [output tank] doesn't go to the exact numbers, so that one was hard." –5th grade student*

Conclusion

- The majority of students said that they enjoyed playing the game, and that it was challenging and fun. Nearly all students would recommend it to a friend, especially their friends who enjoy math. Several students mentioned that the game was much more fun than other math games they have had to play in school in the past, especially during distance learning.
    - *“I would give [the developers] a fist bump.” - 6th grade student*
    - *“Your game is really great, I think everybody can learn from it and the mixture of fractions, percentages, and decimals really help.” - 6th grade student*
    - *“I love the game, the animal design, and the background. It’s cool that they have the bubble design, the grape soda, and everything. I liked everything, except for that animal was staring at me through the computer screen, and I wish there were directions, but other than that I liked everything.” - 5th grade student*

4. RESULT

We developed a new prototype component of our *Tanks* app and its digital manipulative to assess student progress toward mastery of proportional reasoning. The prototype was validated as an assessment against proportional reasoning questions from a large, established standardized test. A pilot study with 238 students in grades five and six, found a significant positive correlation (.52), showing that the prototype (1) assessed the intended mathematical skills, measuring student progress toward mastery, and (2) provides meaningful assessments that are predictive of results in the standardized test. The prototype thus functioned as planned, and students were engaged during gameplay. The results show this prototype can be built out to a reliable assessment tool that can complement existing standardized test, yielding performance information that such tests typically miss.



.52 correlation, based upon 132 observations