Exploring Methods for Tracking Students' Performance through Curriculum Embedded Assessments Designed to Inform and Accelerate Learning

Melissa Hogan, Yigal Rosen, Ilia Rushkin, Barbara Hubert, Maneeza Dawood, Sara Bakken

BrainPOP, New York, NY 10010 USA

Published: December 6, 2023

Abstract

This paper explores new methods for tracking learning growth using curriculum-embedded assessments. The key objective of the study was to examine new ways to capture students’ performance data to help make ecologically-valid inferences about what learners know and can do long before benchmark or end-of-year assessments. Student learning outcomes are examined through the use of embedded assessments and learning growth mapping. A study of 450,934 students finds that continuous engagement with essential skills on an educational technology platform, BrainPOP, is strongly associated with statistically significant increases in learning gains in Literacy/English Language Arts, Math and Science. Educators might consider using embedded assessments to gain access to and provide early feedback for students. By identifying gaps in students' knowledge, teachers can offer timely support to their most vulnerable students.

Introduction

Whereas formal education systems are increasingly exploring ways to engage students to improve learning outcomes, the assessment of learning continues to rely heavily on summative assessments, which are time-consuming, stressful, and point-in-time estimates of learning. Researchers and practitioners have acknowledged the need for assessments that go beyond traditional, large-scale summative assessments and inform instruction: assessments *for* learning (AFL). Educational technology has the power to change the way we think about assessment, particularly in the ways we utilize summative and formative assessments to address learning. Tools to assess learning are of particular importance now in a time of increased concern over current formal education systems that prioritize large-scale assessments (Zheng, Fancsali, Ritter & Berman, 2019; Evans & Lyons, 2017).

In recent years, the use and impact of learning analytics has grown considerably within the field of education. The use and refinement of digital experience platforms has also grown, particularly the need to show concrete learning progression. A significant priority for K-12 education is bridging the gap with digital experiences and enabling a better way to use digital platforms to enhance student learning. One factor driving the growth in showing learning progression is the need for districts to demonstrate improvement on standardized tests in order to receive funding (Baker, Gasevich & Karumbaiah, 2021; Lingard & Lewis, 2016). As such, an important area of research within the field of learning analytics is developing new learning tools that help identify what works for what students and under what circumstances.

This paper uses an assessment for learning approach (AFL) to examine the application of embedded assessments– an integrated system for assessing, interpreting, and monitoring student performance– as well as a robust knowledge tracing algorithm using computational psychometric methodology, through which both students and educators can evaluate learning progression and proficiency long before benchmarking assessments and year-end exams. Furthermore, this paper examines how consistent and sustained student engagement with embedded assessments and learning growth mapping promotes more positive learning outcomes compared to students who have less consistent and sustained engagement with such tools.

Assessment For Learning

Assessment is a vast and varied topic, and is integral for both teaching and learning. Assessment research consists of two primary ideologies: assessment *of* learning and assessment *for* learning (AFL). Assessments of learning primarily include summative assessments such as benchmark assessments, unit tests, end-of-year examinations, and other types of formal testing. Assessments for learning (AFL), on the other hand, include any activities that demonstrate how well a student is learning, and assessing students' learning. Black and WIlliam (1998) identified two sequenced actions that could make assessments formative: learners' awareness of the gap in their current knowledge relative to their learning goals, and the action needed to close that gap (Brown, 2019).

Embedded Assessments and Computational Psychometrics

Developments in both technology and learning science have advanced the world of educational assessment and the need for expanding psychometrics to capture and understand emerging data trends. Embedded assessments are an integrated system for assessing, interpreting, and monitoring student performance (Wilson & Sloane, 2000). Embedded refers to opportunities to assess student progress and performance that are integrated into instructional materials and virtually indistinguishable from day-to-day classroom activities. Educational technologies provide a key opportunity for embedded assessment, with instructional units including a robust system of embedded assessment design to provide teachers with actionable student performance data long before end-of-unit or benchmark exams. Technology-enhanced embedded assessments present students with complex, life-like situations in which they can pursue a sustained investigation. These assessments have the potential to provide diagnostic information not just on the questions students can answer, but also on their capacity to engage in important metacognitive processes to help interpreting performance scores and inform interventions without the interruption of a series of questions and answers (Ercikan, & Pellegrino, 2017). For instance, response processes might include the strategies and approaches that students choose to solve a problem, their motivation and engagement with the item.

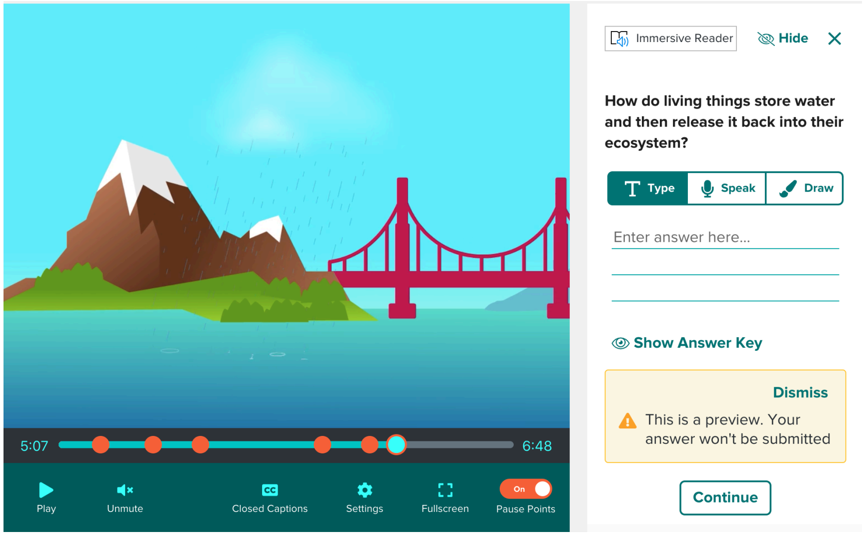
There is an urgent need to expand the methodologies in existing psychometrics to accommodate the challenges from new forms of learning and emerging educational technologies. *Computational psychometrics* refers to the blend of computational techniques and traditional psychometrics to address psychometric questions arising from the changing landscape of assessment. It provides a method to analyze large-scale/high-dimensional learning and provide actionable and meaningful insight based on measurement of individual differences as they pertain to specific skills (Davier et al., 2021). Computational psychometrics can be embedded within a personalized learning experience to facilitate an adaptive experience in which the student is placed in a position to seek out assessment opportunities, rather than have them be imposed by someone else.

BrainPOP Embedded Assessments

BrainPOP (grades 3-8) and BrainPOP Jr. (grades K-3) are learning tools dedicated to helping kids understand their world. BrainPOP's framework successfully links learning and assessment by managing the relationship of learners' data to content (instructional and assessment), essential skills, and analytics. The digital learning platform utilizes three different types of embedded assessments: Pause Points, Quizzes, and Challenges.

Pause Points are learning activities embedded into BrainPOP movies (see figure 1). With Pause Points enabled, a BrainPOP movie will periodically stop, and a prompt will appear on the screen. The user must respond to the prompt in order to continue watching the movie. Pause Points aim to bring to the surface grade-level concepts, support students' understanding of grade-level vocabulary, maintain student engagement with the movie, provide an equitable viewing experience for students watching the movie, and provide an opportunity to practice critical thinking skills. The number of Pause Points varies depending on the length of the movie within that topic. In this study, items with 1 Pause Point per movie minute (4 minutes = 4 Pause Points) and a mix of question types for each Pause Point set were used.

Figure 1. Pause Points embedded assessment.

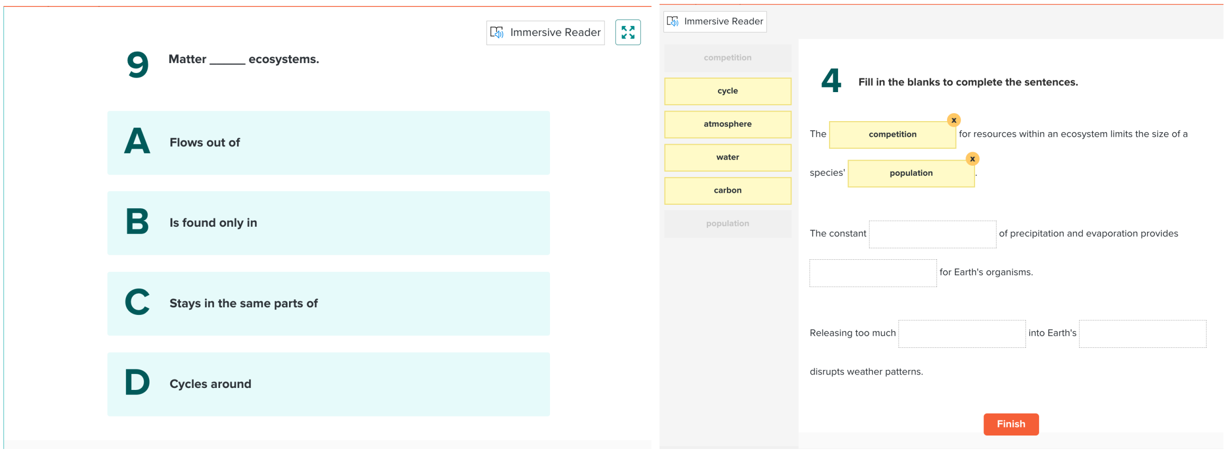

How do living things store water and then release it back into their ecosystem?

Enter answer here...

This is a preview. Your answer won't be submitted

The second type of embedded assessment are Quizzes, which examine the extent to which students understood the main concepts in the movie. Each quiz consists of 10 questions that each contain four answer choices and utilize three different question formats as deemed appropriate for the content in each question. Challenges are the third type of embedded assessment and are typically used to prompt higher-order thinking. Each challenge consists of 4 questions and use question types that are more interactive to heighten cognitive load. Challenge question types may include multiple select, label the diagram, sequencing, matching, venn diagrams, concept maps and text highlighting.

Figure 2. Example of Quiz and Challenge embedded assessments.



9 Matter _____ ecosystems.

A Flows out of

B Is found only in

C Stays in the same parts of

D Cycles around

competition
cycle
atmosphere
water
carbon
population

4 Fill in the blanks to complete the sentences.

The competition for resources within an ecosystem limits the size of a species' population.

The constant ____ of precipitation and evaporation provides ____ for Earth's organisms.

Releasing too much ____ into Earth's ____ disrupts weather patterns.

Although the three types of embedded assessments differ from one another, these learning activities are all designed to assess key grade-level concepts, as well as students' learning progression and proficiency. Inferences are sprinkled throughout the assessments, providing ecological value and relevant content and topics for students. Validation of the three types of embedded assessments using Item Response Theory (IRT) can be found in Appendix A.

Assessing Learner Growth

The Learner Growth Engine utilizes computational psychometrics to construct proficiency from user interactions. It makes use of static data that describes the set of skills, the items and the item-skill tagging - data that indicates which item is aligned with which skills. An example of skills: Common Core State Standards literacy standards. User interactions can be assessment items (quizzes, challenges and pause points) and non-assessment interactions (movies and readings).

In simpler terms, when a student engages with a movie, reading, quiz, challenge or pause point on BrainPOP that is tagged to an essential skill, it records the interaction and produces a proficiency score for that skill.

The tool is different from most applications in that it is capable of using items aligned to multiple skills; It takes into account the “connection strength” between an item and a skill (in particular, if an item is tagged with multiple skills, they need not be all equally strong); it takes into account the depth and the difficulty[1] of items. It incorporates non-assessment items (e.g., movies); allows easy tracing of mastery on skill combinations (multidimensional skills of the Next Generation Science Standards type); allows tracing of skills on different levels of granularity, by exploiting the hierarchical structure of skills if it is provided; produces not just the mastery values but also confidence intervals. The mastery estimation method is built on the regression of scores from assessment items with weights exponentially decaying with time. The skill-specific parameters of that exponential decay, as well as the skill-specific transfer of learning and the initial mastery levels and their weights are trained by predicting the scores in the historical data.[2]

The table of archival mastery that algorithm produces is intended for data reporting: it is the full record of mastery values for combinations student-skill-timepoint, where timepoints are timestamps of all events when the student interacted with an item tagged

[1] In this model, depth influences the outcome symmetrically, while difficulty - anti-symmetrically with respect to the scores received. A deep item has a large evidential weight in calculating mastery, whether the student received a high or a low score, or even if the item was not an assessment item. A difficult item has a large evidential weight if the received score was high, and a small one if the score was low. Non-assessment items do not have difficulty. We believe that the depth property in the model is correlated to the depth of knowledge from cognitive science, however we leave aside the question of the details of their relationship.

[2] The knowledge tracing system ("the learner growth engine") is patented: I. Rushkin, Y. Rosen, US patent # 11,568,753.

with the skill, and thus the mastery value was updated properly aggregated by time and averaged across a group of students (e.g., all students from a certain school or a school district) allows creating “learning curves,” i.e., displaying how the average mastery level of a school on various skills has been changing in time.

In BrainPOP, the tool is set up without real-time capability. The outputs, in the form of proficiency progress visuals for schools and school districts, are available in an out-of-product app.

The key opportunities towards adaptive learning

Since the Learner Growth Engine outputs a data-based estimate of what skills a student knows and how well, it will serve as a data source for informing adaptive learning. We will use it to identify which skills need remediation for a given student, as well as for discovering which items are the most effective for increasing a skill mastery.

Creating the adaptive learning will require the development of the recommendation engine, which will, along with the algorithm outputs, take into account the students’ recent history of interactions and, most likely, implement a form of collaborative filtering (which, in simple terms, means identifying other students who are in some important sense similar to the current one and looking what they did next when they were in a similar situation, i.e. had a similar recent history and a similar knowledge profile). The recommendation calculation will be hosted in an AWS Lambda-function, reacting to an incoming activity signal, waking up to create the recommendation, updating the data and going back to sleep. This will require further extending the architecture that was created for the Learner Growth Engine, using similar solutions.

The Present Study

The goal of this study was to examine comprehensively the effect of students' engagement with essential skills in Literacy/ELA, Math and Science in BrainPOP and BrainPOP Jr. on learning outcomes as measured by validated assessments organically embedded into the learning experiences.

The study was designed to examine empirically the following research questions: (1) What are the effects of students' engagement with essential skills on BrainPOP on their development of proficiency in Literacy/ELA, Math and Science within 90 days of BrainPOP implementation in their classroom, as measured by embedded validated assessments, and (2) What are the effects of students' engagement with essential skills on BrainPOP Jr. on their development of proficiency in Literacy/ELA within 90 days of BrainPOP implementation in their classroom, as measured by embedded validated assessments? We hypothesized that greater engagement with both BrainPOP and BrainPOP Jr. would predict higher scores on the validated embedded assessments.

Implementation Model

We know students build toward skill proficiency with repeated opportunities for engagement and practice. A research-based design and developmental sequence of those opportunities are equally important. Providing support, or scaffolding is a critical component in building new skills. Applebee and Langer (1983) identified key features of instructional scaffolding that include intentionality, appropriateness and structure (i.e., modeling and questioning activities are structured around a model of appropriate approaches to the task and lead to a natural sequence of thought and language). In addition, frequent low-stakes embedded assessment of skills and concepts at strategically spaced intervals can improve how well students retrieve information (Karpicke & Bauernschmidt, 2011). Strategies like interleaving, or studying different problem sets, support identifying similarities between different ideas and concepts to improve long-term learning (Taylor & Rohrer, 2010).

This body of research guides the design of BrainPOP learning activities and implementation recommendations via a learning arc. Each part of the arc focuses on an element of the learning experience and makes explicit how students can engage in the development of skills over time while building content knowledge.The learning arc includes: (1) building knowledge [with movies and pause points], (2) applying and assessing [with quizzes and Challenges], and (3) deepening and extending [with creative tools and learning activities].

BrainPOP recommends that teachers assign learning activities to build knowledge and apply and assess learning at least two times a week for 15 minutes continuously over 90 days in order to see growth in skill proficiency.

Method

This study used an observational, quasi-experimental design, participants were not assigned to treatment and control groups by the research team. Rather, two groups were formed using the student engagement implementation criteria described above. Two different methodologies were used to compare the high engagement group (at least two times a week for 15 minutes continuously over 90 days) to the low engagement group (less than two times a week for 15 minutes continuously over 90 days). All statistical analyses were performed in R. In all cases, Cohen’s *d* was used as a measure of effect size. A Welch’s t-test was used to correct the degrees of freedom for unequal sample sizes and variance between the high and low engagement groups. The first method (Method A- detailed in the results section) averaged learning gains across all users for whom data was available. The second method (Method B- detailed in the results section) offered a more conservative estimate of average change in learning gains compared to method A. In method B, missing data from students were imputed to imply that no change in proficiency occurred. Both methodologies provide different perspectives on learning outcomes in the study sample.

Participants

Data were collected from 450,934 students who used BrainPOP during a 90-day period (February 1 - May 1, 2022). Data were collected from participants within the United States, and were all BrainPOP users, split by the appropriate product, BrainPOP Jr. (K-2) and BrainPOP (3-8). All data were anonymous and de-identified, and no demographic information was collected from participants.

Assigning students to high and low engagement groups

We considered the students' interactions with Essentials content during an approximately 90-day period (February 1, 2022 to May 1, 2022). The interactions with quizzes, challenges and pause points embedded into BrainPOP movies were examined on the level of a whole quiz, challenge, or pause point set, rather than on the level of separate questions. Each content piece belongs to a topic, and we know the correspondence between topics and the three skill categories (Literacy/ELA, Math, Science)[3]. Thus, we analyzed each student's interactions by product (BrainPOP or Jr.)

[3] This correspondence is practically clear-cut. Very rarely it occurs that a topic corresponds to two categories, but one of them is represented in it as a small minority of content pieces, so we ignore that and choose the majority category. This is a rare and small correction and it will not affect the results of the study.

and by skill category. Interactions in the same (product+skill category), if separated by less than an hour, were merged into a single interaction.

The assignment of students to high and low engagement groups was done on the basis of their interaction frequency, expressed in terms of the average interval between interactions (e.g., frequency “once a day” is 7, because it is an average of 7 days between interactions). The target intervals we identified are shown in table 1 below.

Table 1. Target Intervals

Category	High engagement group	Low engagement group
Literacy/ELA	<=14 days	>=42 days
Math	<=28 days	>=56 days
Science	<=28 days	>=56 days

We started by assigning students to high and low engagement groups. We computed the average interval between interactions for each student, product and skill category in two ways: 1) dividing the total time window (approximately 90 days) by the number of interactions; 2) taking half[4] of the maximum observed interval between interactions[5]. If both these quantities satisfied the target condition for the high engagement group, the student was assigned to that group. If the first of these quantities satisfied the target condition for the low engagement group, the student was assigned to that group. Students that are not assigned to either group are dropped from the study.

[4] The halving is the allowance for interactions within the prescribed period. E.g., if the student is interacting once a week but it can occur arbitrarily early or late in the week, the interval between interactions can be as long as 2 weeks.

[5] For N interactions there are N-1 intervals, but we add to that the 0th interval, which is the half-sum of the time interval from beginning of the time window (Feb 01 2022) to the first interaction, and the time interval from the last interaction to the end of the time window (May 01 2022)

The end result is that for each combination of product (BrainPOP or Jr.) and skill-category (Literacy/ELA, Math, Science), we obtained a list of students in the high engagement group and a list of students in the low engagement group. Note that the same student may have landed in low engagement for one product or skill-category but in high engagement for another.

Method A

All students' proficiency on every skill is known after every interaction. Unlike the calculation in the group assignment above, these interactions are on the level of individual quiz/challenge/pause-point questions, and their correspondence to skills is known exactly on that level. We took data from students in the low and high engagement groups, matching on the student ID and the category of skill.[6]

For each skill, we used interactions only within the study time window (~90 days), starting from the 3rd interaction (the first two interactions of a user with a skill are considered the burn-in period, when it is not yet possible to know the level of proficiency with any certainty). We dropped student-skill instances for which, after these conditions, have only one interaction. For the remaining interactions, we record the first (m_1) and the last (m_2) proficiency values. The outcome variable was defined as the incurred percentage proficiency difference $100(m_2/m_1 - 1)$.

We then aggregated the data from the level of students and skills to the level of product and skill category[7], computing a number of descriptive and inferential statistics (e.g

[6] We do not match on product (BrainPOP or BrainPOP Jr.) at this step, but overlaps are virtually impossible anyway, since content of these products differs in the skills it is aligned to. Moreover, it is not expected that the same student commonly interacts with both products.

[7] Note the nature of this aggregation: the same student may have interacted with many skills, all in the same category or also in different categories. Thus, a single student gives rise to multiple user-skill rows in the data that we are aggregating. Furthermore, it may be that one student interacted with all the skills in the category, while another one only with some.

Cohen's d for the outcome variable per group for the effect size, t-test's p-value for the statistical significance) and several other auxiliary measures (e.g., number of students per group, means and standard deviations of m_1 and m_2 per group).

Method B

In the second method we imputed the missing proficiency data m_1 for a skill as a population average and assumed no change (m_1=m_2). This means that when we average the learning gain across skills, the averaging is done across the same full set of skills for all users, but this comes at the price of an additional assumption. Therefore, the learning gains in the second method typically came out lower, as they are a more conservative estimate of imputed proficiencies that show no increase.

Results

We applied the 2-sided unpaired t-test to the high and low engagement groups, using the incurred proficiency difference as the outcome variable. Accordingly, below are the t-statistics, the effect (the mean difference between the two groups), the standard error (the standard error of the mean difference, used as the denominator in the t-statistic formula), and the p-value. We also computed Cohen's *d* as a standard measure of effect size. The effect sizes do not reflect underlying causal relationships as this research was observational. However, these descriptive effect sizes provide useful information to interpret the difference in students' incurred proficiency. The calculations for Method A and Method B are shown below in Table 2 and Table 3.

Table 2. Method A Results

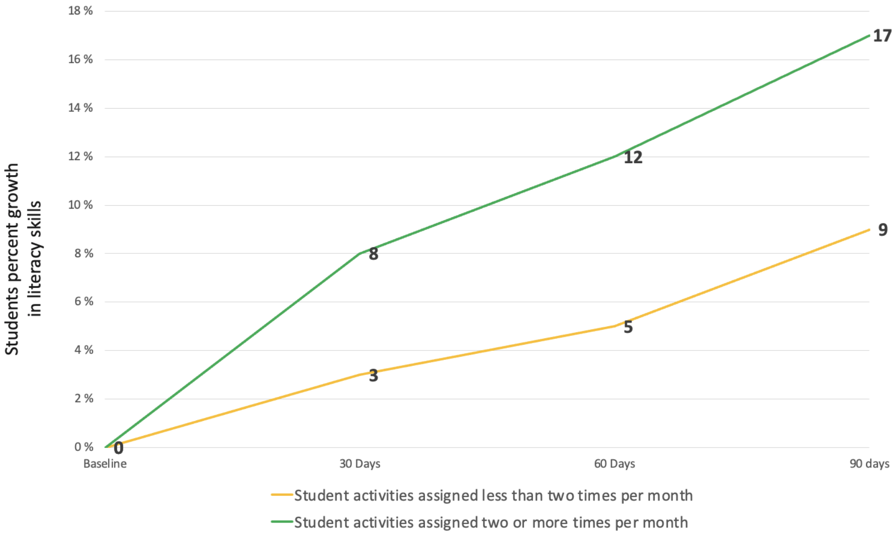
Product	**Category**	**# students in study**	***t*-statistic**	**Effect ± st.error**	**Effect size (Cohen's *d*)**	**p-value**
BrainPOP	Literacy/ELA	285,897	27.0*	(8.0 ± 0.3)%	0.147	p<0.001
	Math	5,288	3.17*	(14 ± 4)%	0.291	p<0.001
	Science	138,294	17.0*	(5.3 ± 0.3)%	0.103	p<0.001
BrainPOP Jr.	ELA	71,567	5.65*	(13 ± 2)%	0.099	p<0.001

Table 3. Method B Results

Product	**Category**	**# students in study**	***t*-statistic**	**Effect ± st.error**	**Effect size (Cohen's *d*)**	**p-value**
BrainPOP	Literacy/ELA	196,275	36.1	(8.0 ± 0.2)%	0.340	p<0.001
	Math	2,989	1.69	(1.9 ± 1.1)%	0.268	p=0.1
	Science	103,880	24.2	(1.97 ± 0.08)%	0.193	p<0.001
BrainPOP Jr.	ELA	61,630	5.39	(9.9 ± 1.8)%	0.170	p<0.001

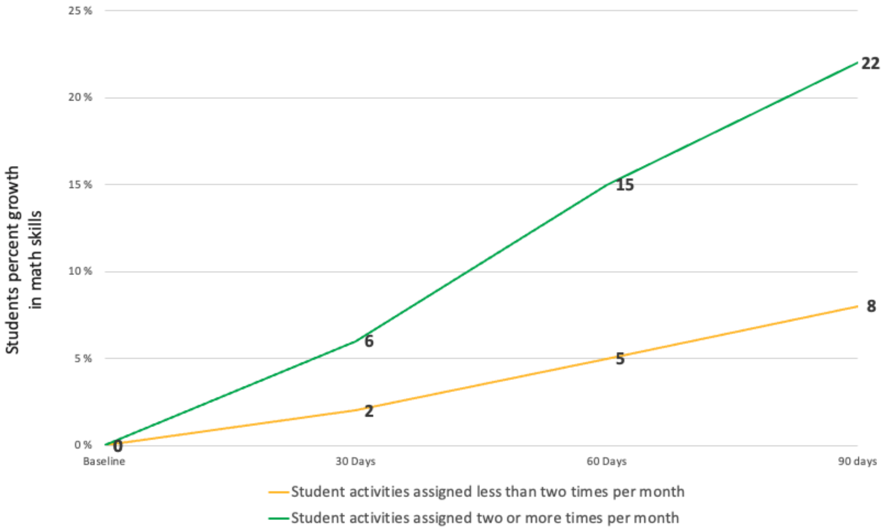
Method A Results

Figure 3. BrainPOP student gains in Literacy/ELA



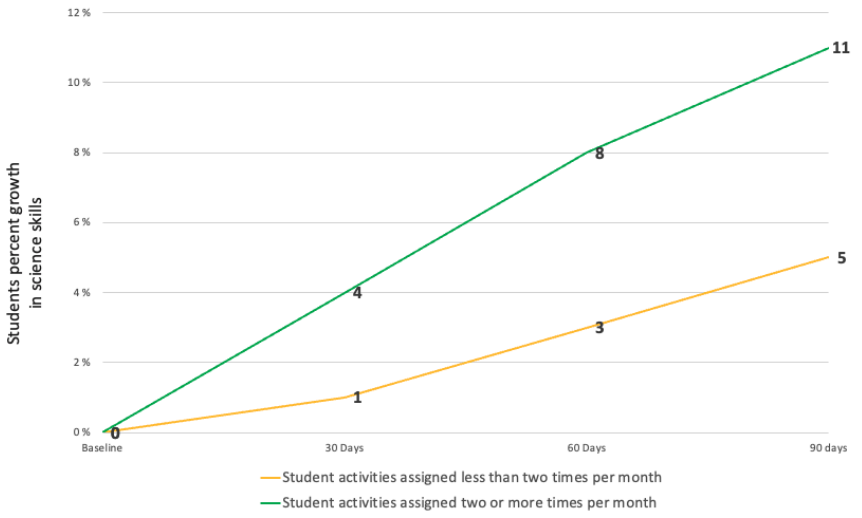
Within 90 days of weekly or once in two weeks engagement with Literacy/ELA skills, BrainPOP leads to 17% increase in learning gains compared to 9% increase for students engaged less than monthly on BrainPOP.

Figure 4. BrainPOP student gains in Math



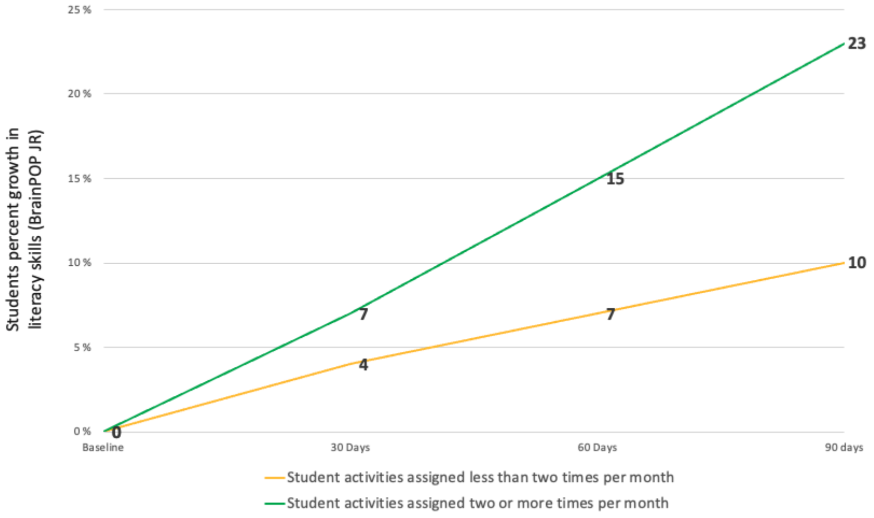
Within 90 days of monthly engagement with BrainPOP Math skills, the high engagement group had a 22% increase in learning gains compared to 8% increase for students who engaged less than monthly on BrainPOP.

Figure 5. BrainPOP student gains in Science



Within 90 days of monthly engagement with Science skills, BrainPOP leads to 11% increase in learning gains compared to 5% increase for students engaged less than monthly on BrainPOP.

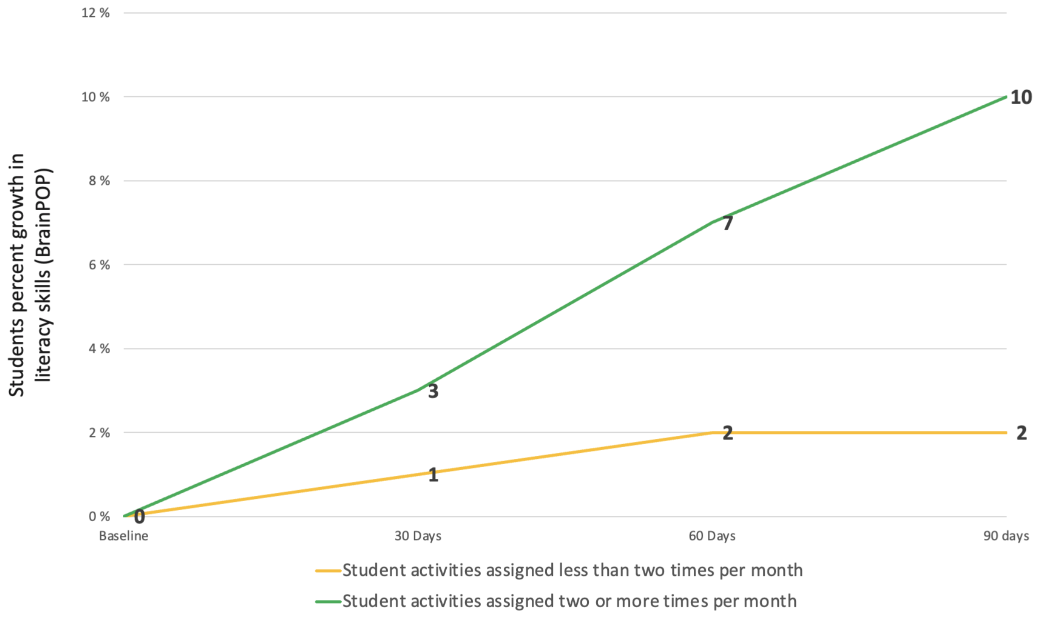
Figure 6. BrainPOP Jr. student gains in Literacy/ELA



Within 90 days of weekly or once in two weeks engagement with Literacy/ELA skills, BrainPOP Jr. leads to 23% increase in learning gains compared to 10% increase for students engaged less than monthly on BrainPOP.

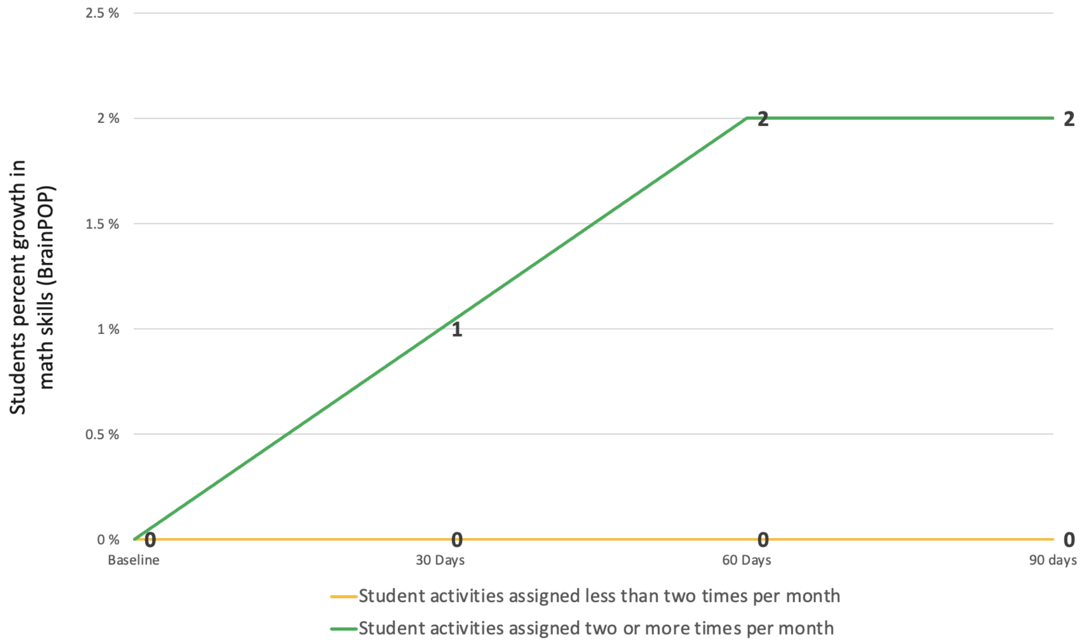
Method B Results

Figure 7. BrainPOP student gains in Literacy/ELA



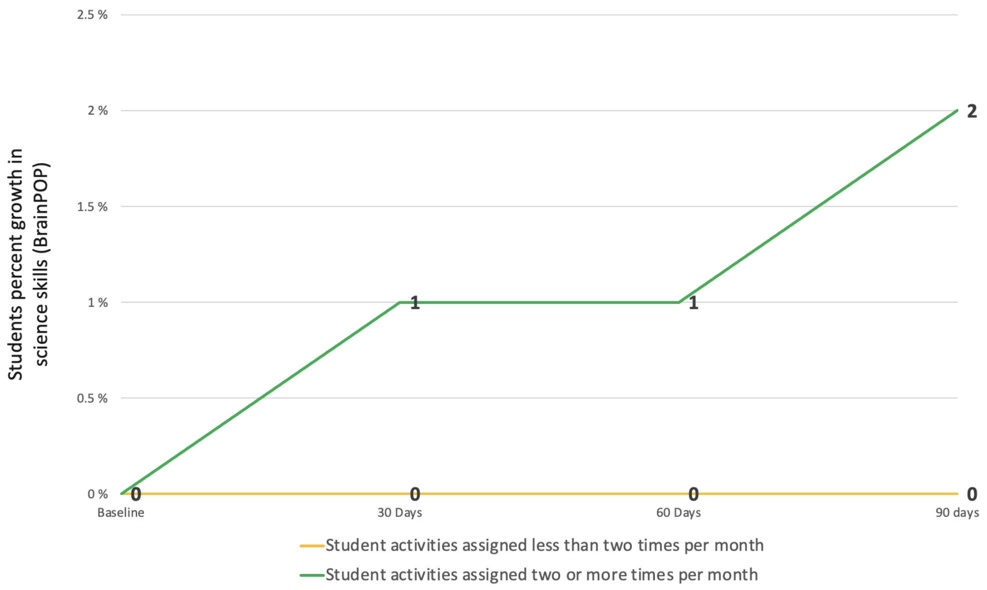
Within 90 days of weekly or once in two weeks engagement with Literacy/ELA skills, BrainPOP leads to 10% increase in learning gains compared to 2% increase for students engaged less than monthly on BrainPOP.

Figure 8. BrainPOP student gains in Math



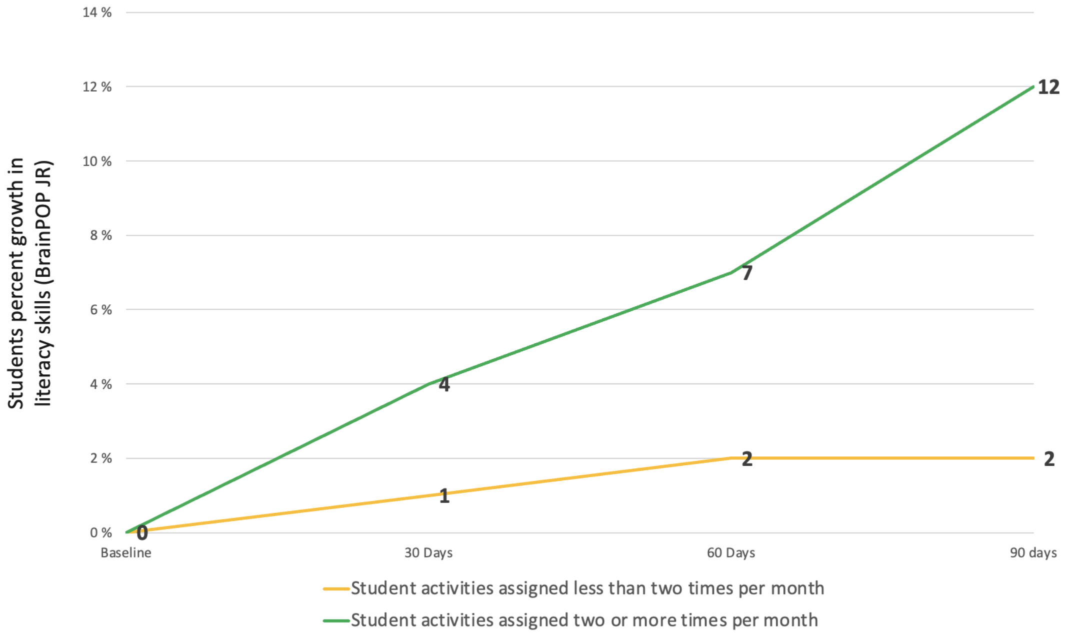
Within 90 days of monthly engagement with BrainPOP Math skills, the high engagement group had a 2% increase in learning gains compared to 0% increase for students who engaged less than monthly on BrainPOP.

Figure 9. BrainPOP student gains in Science



Within 90 days of monthly engagement with Science skills, BrainPOP leads to 2% increase in learning gains compared to 0% increase for students engaged less than monthly on BrainPOP.

Figure 10. BrainPOP Jr. student gains in Literacy/ELA



Within 90 days of weekly or once in two weeks engagement with Literacy/ELA skills, BrainPOP Jr. leads to 12% increase in learning gains compared to 2% increase for students engaged less than monthly on BrainPOP.

Discussion

This paper highlights the utility of BrainPOP embedded assessments and the learner growth engine in assessing student growth over time. The study was conducted for only a 90 day period from February to May, 2022. The level of growth seen is particularly significant when taking into account that it reflects a short window of usage. This study provided preliminary evidence that continuous engagement with essential skills on BrainPOP and BrainPOP Jr. leads to an increase in learning gains in Literacy/ELA, Math and Science skills. In addition, this study demonstrates that students who use BrainPOP and BrainPOP Jr. more frequently make significantly greater gains than students who use BrainPOP and BrainPOP Jr. less frequently.

It is important to note that this research is observational, we examined the natural level of engagement for students who used the product, and students were not assigned to treatment or control groups. We aimed to develop a preliminary understanding of how often students engage with the platform, and how engagement impacts student learning outcomes.

Future studies will be designed to provide further evidence on the correlation between BrainPOP curriculum-embedded assessment and summative assessments such as benchmark and state assessments. In addition, future work will also include continuing our exploration of ways in which administrators and teachers are using reports to assess-for-learning and make data-driven decisions to inform instruction.

Notes for Practice

- Embedded assessments provide early feedback for educators and students, providing insight on the gap between what they know and don't know.
- BrainPOP embedded assessments and learner growth engine provides teachers and students with real-time insight, and shows the progression of their learning over time. A study conducted on over 450,000 students finds that continued use of BrainPOP embedded assessments increases student learning outcomes in ELA, Math, and Science.
- The implications of this research can help teachers reach and help their most vulnerable students earlier.

Declaration of Conflict of Interest

All authors are employed by BrainPOP, which developed the embedded assessments and learner growth engine.

References

Ainsworth, L. (2011). *Rigorous curriculum design: How to create curricular units of study that align standards, instruction, and assessment*. Lead+ Learn Press.

Antonenko, P. D. (2015). The instrumental value of conceptual frameworks in educational technology research. *Educational Technology Research and Development*, *63*(1), 53-71.

Applebee, A. N., & Langer, J. A. (1983). Instructional scaffolding: Reading and writing as natural language activities. *Language arts*, *60*(2), 168-175.

Baker, F.B. (2001). *The Basics of Item Response Theory*. ERIC Clearinghouse on Assessment and Evaluation, University of Maryland College Park, MD.

Baker, R.S., Gasevic, D., Karumbaiah, S. (2021) Four Paradigms in Learning Analytics: Why Paradigm Convergence Matters. *Computers & Education: Artificial Intelligence, 2,* 100021.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7-74.

Bond, M., & Bedenlier, S. (2019). Facilitating student engagement through educational technology: towards a conceptual framework. *Journal of Interactive Media in Education*, *2019*(1).

BrainPOP. (2018). *The Impact of BrainPOP on State Assessment Results: A study of the effectiveness of BrainPOP in grades 3-8*. BrainPOP Research Reports. NY: New York.

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*, 1-29.

Common Core State Standards. (2022). CCSC Standards for math practice. Accessed at http://www.corestandards.org/Math/Practice/.

Ercikan, K., & Pellegrino, J. (2017), *Validation of Score Meaning for the Next Generation of Assessments: The uses of Response Data*. Routledge.

Evans, C. M., & Lyons, S. (2017). Comparability in balanced assessment systems for state accountability. *Educational measurement: issues and practice*, *36*(3), 24-34.

Hubert, B., & Rosen, Y. (2020). *Equity in Learning with BrainPOP: Fostering Access and Impact for All*. BrainPOP Research Reports. NY: New York.

Karpicke, J.D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250-1257. doi:10.1037/a0023436.

Lambert, M., & Sassone, J. (2020). Accelerate, Don't Remediate: An Instructional Framework for Meeting the Needs of the Most Vulnerable Students after COVID School Closures. *Journal for Leadership and Instruction*, *19*(2), 8-13.

Lingard, B., & Lewis, S. (2016). Globalization of the Anglo-American approach to top-down, test-based educational accountability. In *Handbook of Human and Social Conditions in Assessment* (pp. 403-419). Routledge.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

OECD (2021). *21st-Century Readers: Developing Literacy Skills in a Digital World*. PISA, OECD Publishing, Paris.

Rosen, Y. (2009). The effects of an animation-based online learning environment on transfer of knowledge and on motivation for science and technology learning. *Journal of Educational Computing Research, 40*(4), 439-455.

Student Achievement Partners. (2021). 2020-2021 Priority instructional content in ELA/literacy and math. Accessed on 2/29/22 at https://achievethecore.org/content/upload/2020-21%20Priority%20Instructional%20Content%20in%20ELA%20Literacy%20and%20Mathematics_June%202020.pdf

Taylor, K., & Rohrer, D. (2010). The effect of interleaving practice. *Applied Cognitive Psychology, 24*, 837-848. https://doi.org/10.1002/acp.1598.

Wilson, M., & Sloane, K. (2000). From principles to practice: An embedded assessment system. *Applied measurement in education*, *13*(2), 181-208.

Yang, F. M. (2014). Item response theory for measurement validity. *Shanghai archives of Psychiatry*, *26*(3), 171.

Zheng, G., Fancsali, S. E., Ritter, S., & Berman, S. (2019). Using instruction-embedded formative assessment to predict state summative test scores and achievement levels in mathematics. *Journal of Learning Analytics*, *6*(2), 153-174.

Appendix A

Validation of BrainPOP Embedded Assessments Using Item Response Theory (IRT)

The three types of embedded assessments (Pause Points, Quizzes, Challenges) used in the study were validated using Item Response Theory (IRT). IRT is a method of assessing measurement validity. It is a model-based method of estimating parameters for each item included in a scale that separates the individual's responses to the items from the individual's underlying level (or ability) of the latent construct being measured by the scale (Yang, 2014). IRT has become one of the preferred methods of validating scales because it provides a solution for many measurement challenges that exist when constructing a scale.

A multidimensional 2PL item-response-theory approach was applied to the assessment items using the R package "mirt" (Chalmers, 2012). The multidimensionality is dictated by the many-to-many nature of skill tagging. The estimated difficulty and discrimination values of all the items were then mapped to the 0-to-1 range via the cumulative normal distribution Φ (applied to difficulty and to the logarithm of discrimination), which is a standard method of presenting IRT parameters. The distributions of results across the items are shown in Figure 3 and Figure 4, where items are grouped by their type as well as by the category of skills (Literacy, Mathematics, Science).[8] The distribution of difficulties shows that, overall, most assessment items were relatively easy for the students in the study. As expected,

[8] A small fraction of our items are tagged with skills from two categories, and hence such items are included in multiple places on the diagrams. Furthermore, for a small fraction of items all the received scores, after dichotomization, have the same value, which forces these items to drop out from the IRT analysis.

Pause Points were the easiest, followed by Quizzes, then Challenges. Quiz items were more consistently highly-discriminating, whereas pause points contained the most low-discriminating items (Figure 4). Typical benchmarks for interpreting IRT discrimination a parameter are as follows (Baker, 2001):

0-"very low"-0.35-"low"-0.65-"moderate"-1.35-"high"-1.7-"very high".

For the discrimination reported in Figure 4, which is $\Phi(\log a)$, these values become:

0-"very low"-0.15-"low"-0.33-"moderate"-0.62-"high"-0.70-"very high".

These benchmarks are indicated by dashed horizontal lines in Figure 4, where we see that most of our items are of moderate or high discrimination, and some are even in the "very high" category.

Figure 11. IRT Results: Difficulty

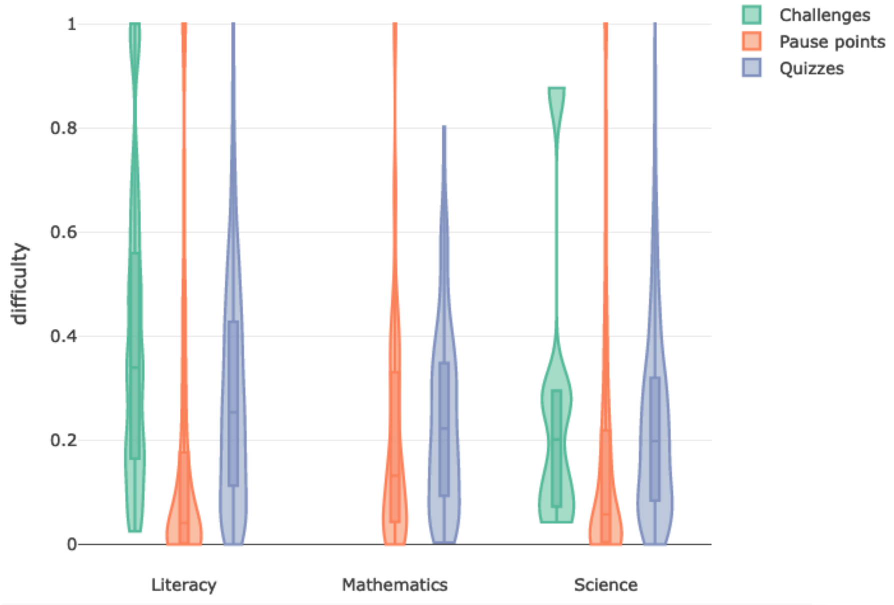
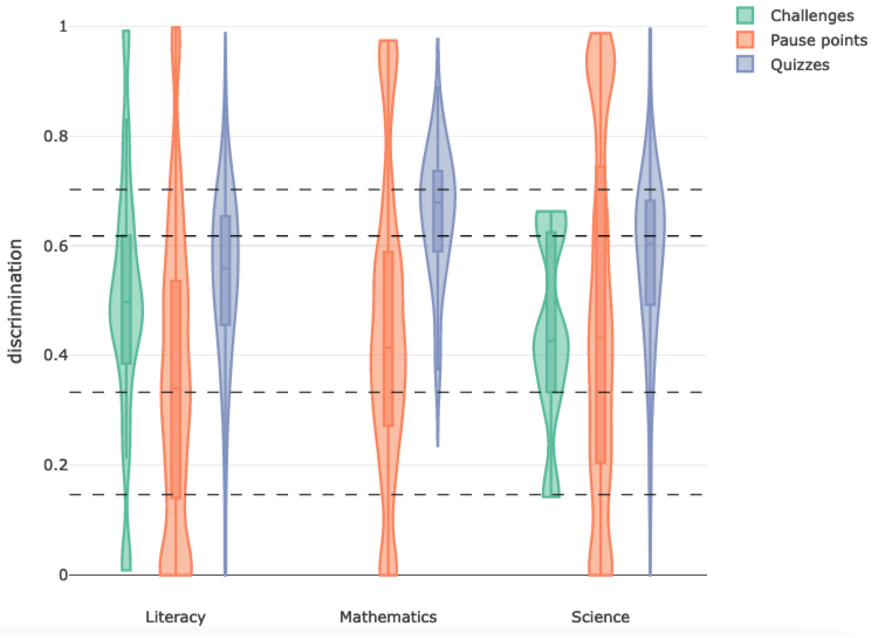


Figure 12. IRT Results: Discrimination. Dashed horizontal lines are the benchmark values for discrimination.


1
0.8
0.6
0.4
0.2
0
discrimination
Literacy
Mathematics
Science
Challenges
Pause points
Quizzes