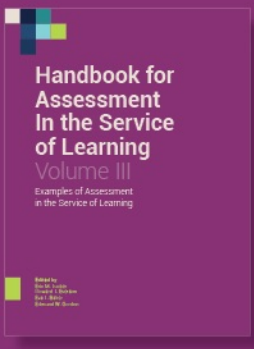
Practical Examples of Assessment in the Service of Learning at PBS KIDS

Jeremy Dane Roberts, Jessica Wise Younger, Kelly Corrado, Cosimo Felline, and Silvia Lovato

UMassAmherst
University Libraries



Series Editors:
Edmund W. Gordon, Stephen G. Sireci, Eleanor Armour-Thomas, Eva L. Baker, Howard T. Everson, and Eric M. Tucker

ISBN: 978-1-945764-33-2

Suggested Citation:
Roberts, J. D., Younger, J. W., Corrado, K., Felline, C., & Lovato, S. (2025). Practical examples of assessment in the service of learning at PBS KIDS. In E. M. Tucker, E. L. Baker, H. T. Everson, & E. W. Gordon (Eds.), *Handbook for assessment in the service of learning, Volume III: Examples of assessment in the service of learning.* University of Massachusetts Amherst Libraries.

Practical Examples of Assessment in the Service of Learning at PBS KIDS

Jeremy Dane Roberts, Jessica Wise Younger, Kelly Corrado, Cosimo Felline, and Silvia Lovato

PBS KIDS, United States

Abstract

This chapter presents several case studies spanning over a decade of work to demonstrate how PBS KIDS integrates assessment in the service of learning to support its mission of providing effective educational experiences at scale. One case study focuses on a video game designed to teach forces and motion, using a dynamic leveling system that adapts to individual player needs. A research study compares this system to a static approach on learning outcomes. Another case study explores how gameplay data is used to assess counting and cardinality skills for players, training neural networks to predict scores on the Test of Early Mathematics Ability. A third case examines the measurement of behavioral changes in gameplay over time across several PBS KIDS games, developing indicators and models to estimate skill development. A fourth case highlights a machine learning competition aimed at understanding the relationship between game/video engagement and performance on interactive assessments in the PBS KIDS Measure Up! app. Lastly, a final case describes using A/B testing to optimize game design variants, balancing engagement and learning to maximize impact. Together, these cases demonstrate the value of assessment in the service of learning at PBS KIDS.

Author Note
The contents of this chapter were developed under a grant from the Department of Education. However, its contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government. [PR/Award No. S295A200004, CFDA No. 84.295A]

Practical Examples of Assessment in the Service of Learning at PBS KIDS

PBS KIDS is committed to making a positive impact on the lives of children through curriculum-based entertainment with positive role models and content designed to nurture a child's total well-being. PBS KIDS' goal is to serve all children. In this chapter, we provide practical examples of how applying the *Principles for Assessment in the Service of Learning* looks in a real-world, scaled up setting. Specifically, we highlight how PBS KIDS, the number one educational media brand for kids (PBS, 2024), has used assessment in the service of learning to further our mission. The work described represents more than a decade of R&D and innovation in learning analytics and learning engineering, driven by our desire to measure, understand, and improve the impact of our media. We have carried out this work in collaboration with a wide range of talented children's media producers, educational researchers, thought leaders, and funders, including the Corporation for Public Broadcasting, the Ready To Learn Program at the U.S. Department of Education, the WGBH Educational Foundation (GBH), University of California, Los Angeles CRESST (UCLA CRESST), and others.

PBS KIDS wants to ensure the media we distribute to millions of children across the US every month (Google Analytics, 2024; Nielsen NPOWER, 2024) have the effect we intend—a positive impact on the lives of all children. In this chapter we focus on how we assess that positive impact through the interactive educational games PBS KIDS distributes. PBS KIDS games offer kids the opportunity to engage with content from a wide range of curriculum in a variety of ways including exploration, tinkering, scaffolded practice, and assessment-focused interactives. These experiences allow kids to explore concepts, practice, get feedback, express what they know, struggle, demonstrate misconceptions, demonstrate mastery, and more. PBS KIDS games present child-relatable situations and challenges, incorporate learning goals, and model problem solving approaches around developmentally appropriate knowledge and skills. The knowledge and skills targeted are selected specifically to help children succeed in school, future work, and life. Accordingly, the

design of the games (and any integrated game-based measurement) incorporates progress, outcomes, and processes, in ways intended to help the children benefit beyond the screens in their everyday life. As needed, PBS KIDS collects fine-grained anonymous user interaction data as children play games to assess different types of learners' knowledge, how it evolves over time, the role our games play in that change, and how we can maximize that role for our media. In this way, we engage in assessment in the service of learning. Children's safety is PBS KIDS' top priority and for that reason, PBS KIDS never collects personally identifiable information.

To date, game-based assessment at PBS KIDS has demonstrated the power of gameplay data to predict scores on standardized tests (Chung et al., 2016; Choi, Suh, Chung, & Redman, 2021), detect (mis)conceptions (Roberts et al., 2019; Lovato, Felline, & Roberts, 2023), assess scientific thinking (Feng, 2019), estimate skill levels for a variety of targeted learning goals including math (Chung et al., 2016), science (Redman et al., 2020; Redman et al., 2021), literacy (Choi, Park, Feng, Redman, & Chung, 2021), and socio-emotional learning (Choi, Suh, Chung, & Redman, 2021), and even measure learning over time (Redman, Feng, Parks, Choi, & Chung, 2023). This chapter will lay out PBS KIDS' vision for assessment in the service of learning in the context of the PBS KIDS mission, audiences, and scale. We provide real-world examples including individualization, assessing skills and measuring impact at scale, and optimizing impact that reflect the following *Principles*:

3. Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.

2. Assessment focus is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.

7. Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.

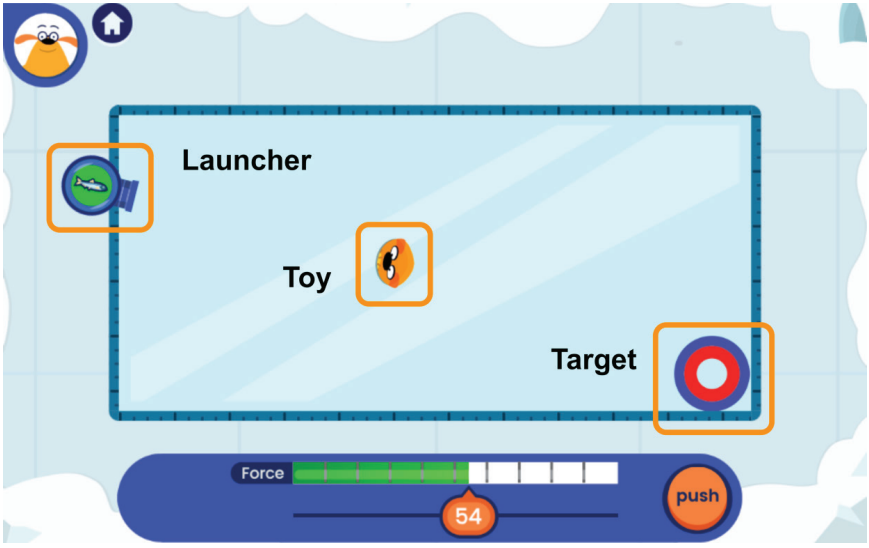
1. Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.

Individualization

PBS KIDS recognizes that not all learners have the same needs. We make great efforts to design content and related measurement properties that work well for as many children as possible. This includes a focus on Universal Design for Learning (a research-based educational framework that guides the development of flexible learning environments and learning spaces that can accommodate individual learning differences; Rose, 2000) to guide design decisions such as avoiding requiring background knowledge, experience, or reading ability that is not necessary. To serve a diverse set of learners requires a diverse set of offerings designed to meet learners where they are. To achieve that, we must assess each player. If we can learn about what an individual knows and doesn't know, what they are struggling with or misunderstanding, then we can use that information to make experiences that respond appropriately and adjust to each individual. PBS KIDS believes that game-based assessment can help power individualized learning experiences, in line with *Principle 3: Assessment design supports learners' processes, such as motivation, attention, engagement, effort, and metacognition.* In the following example, we show how gameplay can be used to estimate a player's skill level and customize their experience by selecting the best next game challenge. Even though this approach did not result in greater learning, it allowed us to understand to what extent a dynamic individualized pathway through a game's levels compares to a static pathway.

As described in Rodriguez, Arena, and Roberts (2018), *Fish Force* is a game that was produced along with videos and activities for the series *The Ruff Ruffman Show* by GBH. *Fish Force* was designed to teach children ages 4–8 concepts of force and motion, like how pushes can have different strengths and cause objects to move in various directions, and how objects can push one another when they touch or collide. Additionally, it was designed to support children in practicing inquiry skills such as making and testing predictions, planning and conducting simple investigations, and engaging in cause-and-effect observations. Players are challenged to rescue a toy plushie stuck on an ice rink by launching a frozen herring at the plushie to knock it onto a target. During the course of the game, players can control the force and/or trajectory of the launcher to attempt to move the plushie to the target area (See Figure 1). Challenge increases between different game levels when additional obstacles are added to the rink—watch out for all of the penguins in the way, ice holes, patches of sand and more! *Fish Force* can be accessed at the PBS KIDS website. (https://pbskids.org/ruff/games/fish-force).

Figure 1.
Example of Fish Force game challenge.



Note. Users can adjust the force meter and the placement of the Launcher to shoot a fish at the Toy to get it to land on the Target while avoiding obstacles. Adapted from Feng, T. (2019). *Using game-based measures to assess children's scientific thinking about force.* [Poster session]. American Educational Research Association Conference, April 5–9, 2019, Toronto, Canada.

In total, 256 game challenges of varying difficulty were created by the *Fish Force* development team, including 128 performance levels (in which the goal is to push the toy to the destination) and 128 prediction levels (of which there are two types: predict the toy's path, or predict where the toy will end up). PBS KIDS games are designed to capture kids' attention, motivate kids to engage deeply, and to be fun so kids invest effort into their play. We theorized we could keep players more deeply engaged by providing them challenges within their zone of proximal development (Vygotsky, 1978). By optimizing engagement, the intent was to promote increased learning outcomes by increasing the amount of instructional material players encountered (Rodriguez, Arena, & Roberts, 2018). That is, rather than provide all learners with the same progression through levels, the game would adapt to support each learner's processes on an individual basis, guiding each player toward the content that would keep them engaged, attentive and motivated, and provide a fun environment to elicit effort to overcome the game's challenges.

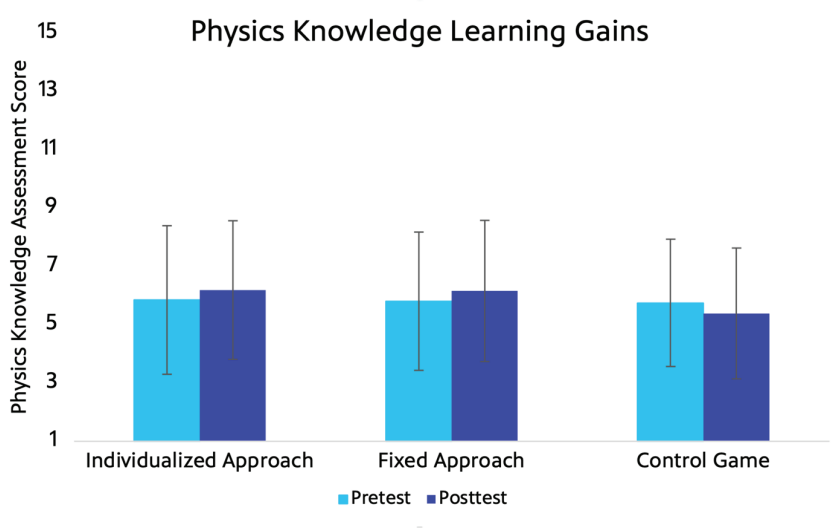
To develop methods for providing an individualized experience, PBS KIDS worked with Kidaptive, a company specialized in individualized learning. Kidaptive applied a Bayesian Item Response Theory (IRT) analysis to rank the difficulty of each game challenge based on an initial sample of players. This model was then incorporated into the game to estimate players' skill levels on the different level types (performance vs prediction) in real time as gameplay proceeded. Similar to computerized-adaptive testing, players' skill estimates were updated after each challenge, and the game used these evolving skill estimates along with the challenge difficulty estimates to select an appropriate next game challenge for the player. Specifically, the probability that the player would correctly solve the next challenge was targeted to be 70%, based on the players' skill level and the challenge difficulty level.

To assess the utility of a personalized approach to gaming, Redman et al. (2019) conducted a study to assess the impact of level progression design on physics knowledge. Students were randomly assigned to play a control game or *Fish Force* with either an individualized level progression (Individualized Approach) or a fixed level progression (Fixed Approach) designed by the game's lead designer and developer. Students were assessed on separate (non-game-embedded) external assessments of children's knowledge of force and motion concepts before and after playing their assigned game. The results (See Figure 2) showed students in both groups that played *Fish Force* made larger learning gains than students

who played a control game. However, the size of the gains was roughly equivalent between the individualized and fixed progression methods.

Figure 2.

Performance on a physics knowledge assessment before and after interacting with the game



Note. Players were assigned to play *Fish Force* with the adaptive level sequence (Individualized Approach), *Fish Force* with the fixed level sequence (Fixed Approach) or a non-physics game (Control Game). The Individualized and Fixed Approach groups showed similar results after controlling for pretest scores. Adapted from Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Schenke, K., Parks, C. B., Michiuye, J. K., Chang, S. M., & Roberts, J. D. (2021). Adaptation evidence from a digital physics game. In H. F. O'Neil, E. L. Baker, R. S. Perez, & S. E. Watson (Eds.), *Using cognitive and affective metrics in educational simulations and games: Applications in school and workplace contexts* (pp. 55–81). Routledge.

This study demonstrated the feasibility of using assessment to support learners' processes, motivation, and engagement in the context of an educational game. The individualized *Fish Force* game that used game-player data to adapt the game in real time was successful in teaching players physics knowledge. However, implementing the individualized approach did not result in significantly greater knowledge gains compared to the static, fixed approach. This finding suggests personalization may not be required for an assessment to be engaging and motivating. It is possible to create effective media children are motivated to engage with without the costs associated with game-specific development to incorporate real-time skill estimation and adaptive leveling.

As a result of these findings, PBS KIDS is now exploring personalization approaches at a broader and ultimately more scalable level. Real-time adjustments to the levels presented to a player within a single game do not necessarily follow *Principle 7. Feedback for the players to clearly address decisions and next step*s. Therefore, instead of personalization within a single game, we are conceptualizing potential approaches to respond to individual needs when selecting items to engage with from the extensive PBS KIDS media library. By incorporating individualization at the library-level (i.e., a recommendation engine), resources could be focused on a small number of strategically representative games that measure skill level for a variety of learning goals. The player-specific information can then be used to guide the overall learning journey for a player. Such an approach would also better align with *Principle 7* by providing clearer next steps for a player to build on their current skill set via the suggested content. This library-level approach may provide higher quality individualization by not only keeping a player engaged at the appropriate challenge level across the media they engage with but also suggesting related or new content to encourage diversifying the topics learned.

Assessing Skills at Scale

In addition to assessing individuals, we believe assessment of our audience has the potential to answer important questions about young children at the group level. The millions of monthly users PBS KIDS games reach (average of 3.4 million unique monthly users on the PBS KIDS Games app and 6.9 million on pbskids.org; Google Analytics, 2024) represents a sizable sample of children aged 2–8, a population that has historically been expensive and difficult to measure systematically, particularly in naturalistic settings (Nagle, Gagnon, & Kidder-Ashley, 2020). As such,

it has been difficult to assess what young children know (their prior knowledge) to understand their educational needs. For example, what are children's skill levels across various subjects, where are their needs greatest, and what are the implications for investments in new educational content? While the United States tracks such information starting in 4th grade through the National Assessment of Educational Progress, no such program exists for preschool, in part due to the difficulty of assessing children this age at scale. This lack of insight into young children's knowledge represents a gap in understanding of kindergarten readiness and the resources needed to support our youngest learners. PBS KIDS believes by designing game-based assessments that meet *Principle 2. Assessment that focuses on progress, outcomes, and processes that can be transferred to other settings, situations, and conditions,* we can inform PBS KIDS' curriculum focus over time to meet demonstrated needs in particular areas. For example, if a particular skill set sees a dip in performance, PBS KIDS can adjust production to develop more related media or better promote and make more discoverable existing content that responds to the need. Below, we discuss an example that demonstrates a proof of concept for such population-level assessment of children's knowledge via gameplay data.

Curious George Busy Day is a set of 16 games, available in English and Spanish, that were developed by GBH, and that focus on counting and cardinality. The set of 16 games represent learning goals such as number knowledge and counting skill. Three games, *Apple Picking*, *Blast Off*, and *Meatball Launcher*, have game mechanics that require players to make a judgment about numbers and actions and therefore can be used to assess player skill level. Specifically, *Apple Picking* assesses a player's ability to count on by ones from a number other than 1 by requiring players to select the missing number in a sequence. *Blast Off* assesses the ability to count backwards from 10 by asking players to select a series of numbers from largest to smallest. Finally, *Meatball Launcher* assesses the ability to count or put out 1 to 5 objects upon request by asking players to give a requested number of items. These tasks are illustrated in Figure 3 and can be accessed at the PBS KIDS website. (https://pbskids.org/curiousgeorge/busyday).

Figure 3.
Example game challenges from Curious George Busy Day

Note. In Apple Picking (A) players must select the apple with the number that belongs where the question mark is in the line of apples. In Blast Off (B), players must select the numbers from largest to smallest to blast off the rocket. In Meatball Launcher (C), players must put the requested number of meatballs on the plate.

As described in Roberts et al. (2018), researchers at UCLA CRESST first conducted analyses examining whether measures of game progress (rounds completed, time spent, time to correct answer), game performance (number of correct first attempts, number of overall correct attempts, number of overall incorrect attempts), or their combination were related to scores on a standardized assessment, the Test of Early Mathematics Ability, 3rd Edition (TEMA-3; Ginsburg & Baroody, 2003). Generally, performance-based measures were more strongly related to test scores than progress-based measures. Across all three games, the strongest positive predictor of math knowledge was the number of correct first attempts at a solution, while the strongest negative predictor was the number of incorrect solution attempts. However, measures that incorporated both progress and performance yielded the highest correlations with the TEMA-3. Specifically, vector combinations that incorporated success in one dimension, but error in another; number of first correct attempts (success) and time taken to correct first attempt (error) ranged from 0.43 to 0.58 and number of incorrect attempts (error) and highest level reached (success) ranged from 0.48 to 0.76 across all three games (See Table 1). Interestingly, *Meatball Launcher* consistently had strong correlations with TEMA-3 scores across all measures. This game was the only examined game that did not provide feedback as to the accuracy of the answer. This finding suggests children do incorporate in-game feedback into their gameplay and can learn from the test. For PBS KIDS, the implication is that if we wish to assess our audience's skill level, some games should be designed solely for assessment purposes (not a hybrid of instruction and assessment) to provide a more accurate measurement.

Table 1.

Correlations (Spearman) Between Vector-Based Angular Component Measures and TEMA-3 Measures by Game

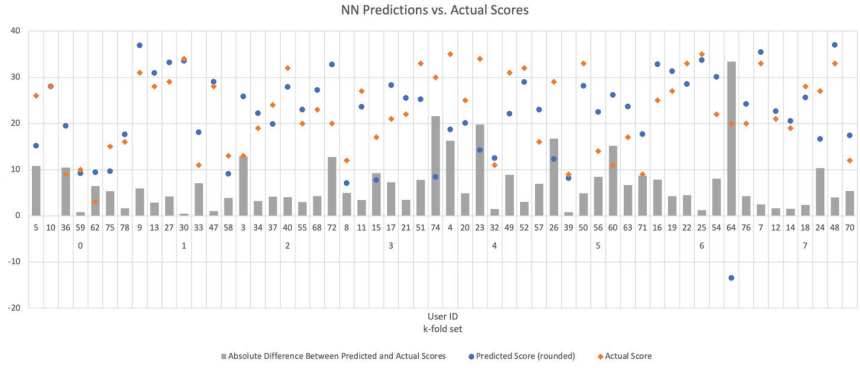
Measure	Total score	Cardinality subscale	Counting subscale
Vector 1: y = No. of correct first attempts, x = Time taken for first attempts (min.)			
Apple Picking	.43**	.37**	.35**
Blast Off	.51***	.40**	.41**
Meatball Launcher	.58***	.52***	.50***
Vector 2: y = No. of correct attempts, x = Mean level time (min.)			
Apple Picking	.26	.20	.12
Blast Off	.42**	.34*	.30*
Meatball Launcher	.70***	.63***	.61***
Vector 3: y = No. of incorrect attempts, x = Mean level time (min.)			
Apple Picking	-.28*	-.26	-.32*
Blast Off	-.28	-.19	-.23
Meatball Launcher	-.38*	-.40*	-.34*
Vector 4: y = No. of correct attempts, x = Highest level reached			
Apple Picking	-.35*	-.34*	-.23
Blast Off	-.48***	-.43**	-.40**
Meatball Launcher	.09	.13	.13
Vector 5: y = No. of incorrect attempts, x = Highest level reached			
Apple Picking	-.48***	-.41**	-.36**
Blast Off	-.57***	-.52***	-.51***
Meatball Launcher	-.76***	-.71***	-.64***

*$p < .05$ (two-tailed). **$p < .01$ (two-tailed). ***$p < .001$ (two-tailed).

Note: From Chung, G. K. W. K., & Parks, C. (2015b). *Bundle 1 computational model analysis report* (Deliverable to PBS KIDS). University of California, National Center for Research on Evaluation, Standards, and Student Testing.

UCLA CRESST then examined how more game-based information about a player might be used to improve predictions on their standardized test performance (Chung & Parks, 2015b). Using 1702 different indicators derived from seven different games from *Curious George Busy Day*, UCLA CRESST built and trained several neural network models. The models were each trained on one subset of data, then validated on another subset. The best-performing model that leveraged data from many indicators of skill across seven games on average predicted individual's TEMA-3 scores within about 8% of their actual score (See Figure 4).

Figure 4.
Neural Net (NN) TEMA-3 predicted and actual scores



Note: Adapted from Roberts, J. D., Parks, C. B., Chung, G. K. W. K., Redman, E. J. K., Schenke, K., & Felline, C. (2018). Innovations in evidence and analysis: The PBS KIDS Learning Analytics Platform and the research it supports. In *Getting Ready to Learn* (pp. 231–248). Routledge.

The results of this study demonstrated the very real potential to use games as assessments that meet *Principle 2: Assessment focus is explicit and includes purposes, outcomes, progress indicators, and processes that can be transferred to other settings, situations, and conditions.* Performance on the selected *Curious George Busy Day* games relates to a completely different and meaningful context: performance on the TEMA-3 standardized test. This work shows that such assessment can be done at scale with young children and demonstrates PBS KIDS is capable of performing benchmarking at the population level through our games.

Measuring Impact at Scale

PBS KIDS serves the American public at scale, and desires to measure the impact we make with media at scale. We define impact as a combination of reach, engagement, and learning effectiveness. For a child to learn something from PBS KIDS media, we must reach them, they must choose to engage, and the media must be effective at promoting learning. PBS KIDS believes that in-game assessments can help us measure the learning component of our impact by following *Principle 7: Assessment quality and validity should be available and reflect evidence related to assessment purpose to permit appropriate inferences and findings about quality, utility, and credibility.*

Measuring reach (number of users exposed to our content) and engagement (how long a user engages, amount of content engaged with, etc.) are relatively straightforward. Measuring learning effectiveness is much more difficult. Unlike reach and engagement, which can largely be measured by counting users and their interactions, effectiveness implies measuring a change in users' performance over a period of time. Historically, large scale randomized control trial (RCT) studies have provided such information on PBS KIDS media. However, these efforts have limitations, particularly when considering employing them at scale. While still considered the 'gold standard' for determining the instructional potential of specific pieces of media, these RCTs are slow and expensive, and do not always reflect how the content is used "in the wild" (Redman et al., 2021). These limitations result in RCTs being conducted on only a small subset of content and leave the effectiveness of the media when used under typical, unguided conditions unclear. Specifically, RCT participants are often directed to use the material in a prescribed, consistent way over a period of weeks, and the material is often in isolation from any other PBS KIDS offerings. However, in non-research settings, users interact with the same game content from within a much larger suite of media offerings (as of 2024, the PBS KIDS Games app offers almost 300 games), and engagement patterns can differ substantially. For one studied set of games, less than 1% of the PBS KIDS Games app population engaged with the games to a similar depth of content coverage as the recruited study population (Choi, Suh, Chung, & Redman, 2021), signaling a potential lack of effectiveness for our population. However, this comparison between populations did find support for the generalizability of efficacy of our content for our population, as gameplay performance and the skill level estimates from psychometric models were similar between the recruited study sample and those players that engaged at a similar level in natural settings.

In the hopes of measuring learning effectiveness faster, more cost effectively and with a more naturalistic sample, PBS KIDS and UCLA CRESST set out to develop a way to use gameplay data from the PBS KIDS audience to directly measure changes in behaviors that are consistent with a player learning over time. Further, to maintain children's privacy, this work had to be conducted with anonymous gameplay data and not incorporate any demographic information. Such an endeavor would extend previous work aimed at assessing an individual's skill at a single point in time (e.g., Roberts et al., 2018; Roberts et al., 2019) to follow how that individual's skill differed across multiple timepoints. While a logical and relatively straightforward extension of previous work, this project presented new challenges. Specifically, we sought to understand whether the changes in behavior could be reasonably attributed to a player's interaction with the PBS KIDS game without knowledge of activities done outside of their interactions with PBS KIDS games. However, if successful, the work could be used to develop an indicator of learning effectiveness that is consistently monitored and reported on, similar to the metrics used for reach and engagement.

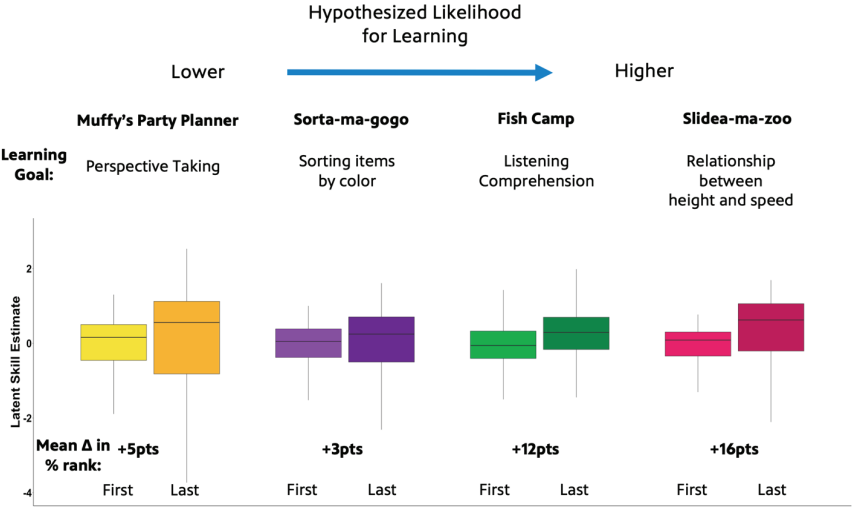
As part of the initial effort at measuring learning over time, Redman et al. (2023) first selected a subset of PBS KIDS games from which skill level at a given construct could be reasonably estimated using gameplay data alone. These games were then evaluated for the potential to promote learning based on features of the games, such as whether user feedback was provided or constructed learning (Nanjappa & Grant, 2003) was encouraged. This evaluation was called a "qualitative ratings validation approach". The four games included in the final analysis represented a range of potential for learning. Specifically, based on the availability and quality of feedback mechanics (incorrect answer elaboration, graduated feedback) and constructive learning processes (prediction, reflection, and debugging/correction), *Slidea-ma-zoo* from the series *The Cat in the Hat Knows a Lot About That!* and *Fish Camp* from *Molly of Denali* were designated as having a high potential for learning. The game *Sorta-ma-gogo* from *The Cat in the Hat Knows a Lot About That!* did not have as much elaborative feedback or encourage player reflection and so was rated as having less potential for learning. Finally, *Muffy's Party Planner* from the series *Arthur* was specifically designed to measure and not teach. Therefore, there was no feedback or constructive processes involved in the game, and it was rated as low potential for learning.

The inclusion of *Muffy's Party Planner* designed for measurement only was key for our validation process. Namely, by examining games with both low- and high-likelihood of learning, we could assess how likely skill gains were due to engagement with the PBS KIDS games. Indeed, young children should be improving their skill sets over time through a variety of opportunities in their daily life, so learning gains not specific to interactions with the game were expected. This qualitative ratings validation approach was determined to be faster and more cost effective than implementing a small, recruited study looking at correlations between external measures and gameplay-based estimate of skill. For PBS KIDS, this work represented a novel symmetric approach to simultaneously validate the utility of game-based performance measures as indicators of skill on a construct, the models used to estimate player skill level, and the qualitative rating system for a game's likelihood of learning. It also provided key data on how much confidence we should have in these approaches.

Next, for each game, UCLA CRESST used an IRT model to estimate the difficulty and discrimination parameters of each challenge or 'item' in a game. A player's skill level on a given construct targeted by the game was then estimated at two time points at least one day apart based on the player's responses to game challenges and the item parameters. Skill change score was determined by subtracting the initial estimate from the second estimate. We found that, as expected, changes in skill were detected across all games. Importantly, though, the size of the gains was generally larger for games with higher potential for learning and lower for games with lower likelihood for learning (See Figure 5). Players of both games rated as having high potential for learning showed larger gains in skill estimate over two time points compared to the changes seen in skill estimates of players of *Muffy's Party Planner*, the measurement game. This initial effort took a conservative approach to provide preliminary proof of concept to measure the efficacy of a game using only anonymous gameplay data. Specifically, strict inclusion criteria, including requiring participants to interact with specific game challenges more than once, and at least one day apart, resulted in only about 10% (N=237,293) of the full data set (N=2,174,787) being analyzable in the model of skill change. Further, while the users included in the analysis showed significantly higher engagement with the games compared to those not analyzed, similar to the comparison between recruited study participants and PBS KIDS Games app users at large, initial performance between the included and excluded players was similar. This comparison indicated that the included players likely did not have greater initial skill compared to the excluded players and suggested the non-studied players would have similar potential for learning gains if they interacted with the game more.

Figure 5.

Latent skill estimates from first and last encounter with a game



Note. Learning gains were roughly consistent with the hypothesized likelihood for learning developed from feature analysis. From Younger, J. W., Roberts, J. D., Felline, C., Corrado, K., & Lovato, S. *The role of learning analytics at PBS KIDS*. [Poster session]. Biennial Meeting of the International Mind Brain and Education Society, July 10–12, 2024, Leuven, BE.

Future work is planned to create models of skill that better reflect the needs of our PBS KIDS audience, namely, constructing valid models for detecting skill change within a shorter period of time that better align with the natural game engagement pattern of our users. Further, we plan to include additional input to the model such as amount of content covered within a session, time between sessions, and more to further refine our models to make better inferences about the effectiveness of PBS KIDS games (Chung, Redman, & Choi, 2023). In this way, as outlined in *Principle 6*, we expect to ensure that our assessment purposes fit our audience, improve the credibility of our assessments, and draw appropriate inferences from them, ultimately helping us succeed in meeting the PBS KIDS mission.

Optimizing Impact at Scale

Beyond measuring impact, another key use case for assessments for PBS KIDS is to continuously improve our approaches toward impact over time. To achieve this goal requires following *Principle 1: Assessment transparency provides clear information about assessment content and use to assist learners, teachers, administrators, and parents.* More specifically, by understanding the specific engagement patterns of individuals as they play games and measuring learning gains as they play, we can determine the relationship between what players do and what they learn. A clear picture of this relationship will help us develop models of impact and improve them over time. However, not all our efforts to link player behavior and learning help move our models of impact forward. In the following examples, we present cases that demonstrate how the assessment transparency noted in *Principle 1* can directly impact our ability to serve learning.

In 2019, PBS KIDS was part of a competition focused on Artificial Intelligence (AI), and its application to various disciplines and hard problems (Felline et al., 2019). Competitors from across the globe tackled a specific challenge and competed on well-defined scoring criteria. The challenge was to use anonymous interaction data from users engaging with a variety of video and games to predict performance on embedded interactive assessments within the PBS KIDS *Measure Up!* app, which was developed to teach preschool and early elementary school-aged children measurement concepts such as height and length, weight, and capacity. The hope was the competitors would help extend previous efforts showing the app was successful in improving children's knowledge of pan balances (Schenke et al., 2020) to understand why children were likely to benefit.

PBS KIDS viewed this data challenge as an opportunity to understand how AI and machine learning approaches could help discover relationships between engagement with various specific features of the media and outcomes on the assessments. Sophisticated models were able to predict players' game-based assessment performance based on their game interaction data reasonably well. Models were scored on a scale of -1 to 1 using methods of measuring inter-rater reliability between the model predicted scores and the actual scores (quadratic weighted kappa; McHugh, 2012). The winning model achieved a score of 0.568, with 0.6 considered a very good score. However, the winning teams employed models that could not be fully explained to PBS KIDS. As such, there was no way to gain insights into the media design choices in the studied games to enable these

predictions to be applied in other scenarios or even iterated on in a theory-driven way. PBS KIDS considers the limitations of such models to be serious enough that we have shifted our focus almost exclusively to explainable models. Without assessment transparency, PBS KIDS cannot improve our models for impact.

PBS KIDS is now taking a different approach to obtain the assessment transparency needed to power the continuous improvement of the design of educational media. We are now conducting randomized control trials directly within a PBS KIDS flagship distribution product: the PBS KIDS Games app. Our approach is to have each experiment-capable game incorporate several variable experiences that can each be independently manipulated. In this way, many different aspects of game design and their potential interactive effects can be examined within the same context of a given game design. Maximizing the experimental space of a given game also allows us to conduct fast analytics-based randomized control trials at the scale of the PBS KIDS Games app audience. At the time of writing, we have completed experiments on two different games that each have the target goal of teaching players about the design process, though designed with different age groups in mind. As explained in Younger et al. (2024), both experiments examined how the level of specificity of game instructions impacted player behavior (Mayer, 2023). The first experiment with 1,054,651 enrolled users additionally examined the effect of prompt construction, comparing a question vs statement format (King, 1991). The second experiment with 567,267 users additionally examined the impact of motivational elements in the game. Through these experiments, we were able to identify which elements of a game are likely to be most impactful to players' experience. In the first experiment, the instruction specificity variable was manipulated within instructions that were verbal in nature (either read by or spoken to the player by in-game). The variable was implemented in two different phases of the game with the intent to determine whether specificity would be more impactful at different phases of learning or whether there may be additive effects (e.g., two specific instructions might be more impactful than one). Yet, there were no meaningful differences across our different experimental conditions. Indeed, as many as 30% of users chose to skip the instructional prompt with the experimental manipulation, though these users did not perform differently from those that did not skip the instruction. We hypothesized multiple explanations for this finding. First, the timing of the specific instructions relative to expected user actions may not have been appropriate to impact user behavior. Second, the presence of additional supportive visual elements present in the game at the time

of instruction were more salient to players than the verbal instructions presented. The transparency of our assessment methods allowed us to iterate on these ideas in future experiments. In the second experiment, we adjusted the manipulation such that it took place in earlier, initial instructions to the user that were visual in nature rather than verbal. In this experiment, the different variable conditions did produce meaningful differences in user behaviors in the game. Those players that received more specific instruction were more likely to make use of features in the game designed to aid performance and required fewer attempts to complete the challenges presented in the game compared with users who received less specific instruction.

While analysis can identify which variants might be most effective for learning, multiple lenses are required to determine the overall impact of a variant. As mentioned earlier, the informal media landscape is filled with many activities for kids to choose to engage with. It is therefore not enough for an experience to be highly educational alone. If kids choose to engage in something else and engagement with our media goes to zero, then impact also goes to zero. Therefore, in addition to comparing how variables influence learning, we consider whether they affect engagement. For example, although we may have chosen to make the instructional prompts non-skippable, through prior work, we know that engagement with a game tends to drop if instructions are required before users can interact with the game. Therefore, while a variable might influence how many attempts it takes a player to solve a particular challenge, we must also ensure players are engaging with the same number of challenges across all experimental conditions. What is the proper balance between engagement and effectiveness? In the experiments run to date, there were no differences in engagement across experimental groups. However, as we expand our experiment program to different types of variables, it is our hope to establish a quantitative understanding of the balance between engagement and effectiveness. Ultimately, this foundation will support team debate, definition, and alignment toward a quantitative definition of impact itself and how impact is aggregated across millions of users and relevant subgroups. As we establish a baseline understanding of what is true today, we will use this understanding to help us improve going forward. Developing the capability to discover the optimal design principles of educational games will provide the feedback that game producers, designers, and developers need to help make decisions about how to proceed with game development iterations, and with future game design efforts.

Progress and Implications

PBS KIDS has been fortunate to develop and execute a variety of projects that all focus on using assessment in the service of learning. This program of work has required over a decade of systematic work across children's media producers, educational researchers, thought leaders, and funders to innovate the tools, technology, and processes needed to measure, understand, and improve PBS KIDS games. First and foremost, the game-based assessment work would not be possible without data collection infrastructure. Over the years, PBS KIDS developed a bespoke system for data collection to meet the many needs for our research program. We capture very detailed anonymous interaction data (which includes no personally identifiable information) from PBS KIDS games. Data collected by our system includes events capturing time series data around user action, system reactions, instruction, feedback, hints, voice over captions, and snapshots of the evolving state of game challenges such as puzzles and problem-solving tasks. As such, much more data are generated from our system compared to more typical business use cases aimed at understanding user activity. Therefore, as we collected more data from more sources across games and distribution channels, we developed tools for great control over when and where data are collected. This high degree of control has the dual benefit of supporting both privacy and sustainability goals. Other important steps to scaling data collection include standardizing log data across games to allow for greater consistency and efficiency of analysis and the game development process itself. For example, PBS KIDS has certain requirements for games distributed on its platform. By fitting our data collection platform into this ecosystem, we could more easily ensure all games commissioned by PBS KIDS have the potential to use our system if desired.

Our approach to data collection leads to interesting limitations in the data collected such as the absence of information about a player's background, demographics, specific setting, and a lack of knowledge of whether a single device is being shared amongst multiple individuals during co-play. Despite these limitations, as the examples above show, the data power research that is safe and valuable. Further, to supplement the large-scale anonymous data collected, PBS KIDS also commissions recruited studies that can collect additional demographic data through formal research consent processes. A series of tools (e.g., to easily deploy games into research environments, configure data collection, and provide researchers with easy access to study data) were created to facilitate these

studies and enable PBS KIDS games to be researched in more controlled settings and ensure research data is separate and distinct from that collected from the general population.

Another equally important contributor to the success of our research program has been the cultivation of data awareness and use of gameplay data, assessment, and the related potential for measuring and optimizing impact. We have strived to amplify the results of our work both internally to product development and strategy teams and externally to academic and industry groups. Meeting these goals has required research agendas that are developed in a mutually beneficial fashion, contributing to both foundational work around the potential to use game-based assessment for learning as well as more immediate tangible benefits to the wider PBS KIDS community. For example, during launches of new games, highly detailed user interaction data are collected with the intent of understanding how to measure learning from player behavior. These same data can be used to understand important player patterns such as where players may encounter unexpected difficulty with the game, which can be reported back to the game developers who can adjust the game as necessary. Building such symbiotic research programs has emphasized the importance of individualizing our approaches to learning and teaching within our own team, and across the community of production partners. Just as we develop different games to meet the needs of different learners, we have had to evolve our research programs to meet the needs of different consumers of our work. Adapting to meet the needs of our consumers has resulted in developing analytic pipelines that can operate on different time scales. An academic pipeline, for example, might take place on a longer time scale and include detailed statistical analysis presented in a formal report. A game development pipeline, on the other hand, may operate on a much faster scale, taking samples of data and using visualizations to quickly assess whether a feature seems to be working or not. This allows data-informed iteration and improvement to be seamlessly integrated into our development processes, which is considered vital to PBS KIDS and our digital producers. Communicating insights in a way that is familiar and approachable for different audiences has been instrumental to growing our support base, and therefore our research program capabilities. There is much more for us to explore around how best to support the collaborations and processes that power the development of PBS KIDS games, distribution platforms, user experiences, marketing and promotional strategies, distribution strategies and more.

Looking forward, we hope to continue our efforts related to assessment in the service of learning on multiple fronts. First, we want to continue to innovate on how we develop and validate new models for assessment. This effort includes continuing to improve how we determine whether models are suitable for the purposes for which we create them. In particular, we want to ensure the inferences we make and the decisions we take based on them are aligned with our objectives. Next, we want to expand our effort to support learner's processes with individualized instruction in ways that encompass the larger PBS KIDS library of media, including both games and videos. We are currently in early exploration and planning around recommendation engines, and how they can be applied appropriately in the PBS KIDS context and expect to learn a lot over the next few years. Finally, we want to further demonstrate that the skills players exhibit while playing PBS KIDS games (such as the *Curious George Busy Day* games) can transfer to other different and important contexts beyond the TEMA-3, e.g., on performance tasks in the real world.

After over a decade of work and a variety of principles coming together, collectively we have accomplished much. We have developed large-scale, high value, and safe gameplay data collection capabilities to power game-based assessment-powered individualized learning approaches, models to estimate skill levels on learning goals using gameplay, and models for estimating learning over time based on the skill estimates. We have further crafted a method for the systematic, speedy, and efficient discovery (and improvement over time) of design principles for educational children's media that work best at scale. What will the next decade bring?

References

Choi, K., Parks, C. B., Feng, T., Redman, E. J. K. H., & Chung, G. K. W. K. (2021). *Molly of Denali Analytics Validation Study Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Choi, K., Suh, Y. S., Chung, G. K. W. K., & Redman, E. J. K. H. (2021). *Population study final study report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Chung, G. K. W. K., & Parks, C. (2015). *Bundle 1 computational model analysis report* (Deliverable to PBS KIDS). University of California, National Center for Research on Evaluation, Standards, and Student Testing.

Chung, G. K. W. K., Parks, C. B., Redman, E. J. K. H., Choi, K., Kim, J., Madni, A., & Baker, E. L. (2016). *PBS KIDS Final Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Chung, G. K. W. K., Redman, E. J. K. H., & Choi, K. (2023). *Wombats Analytics Evaluation–Final Plan* (Final deliverable to PBS KIDS). UCLA/CRESST.

Felline, C., Roberts, J. D., Rohner, J., Oder, J., Springer, K., Corrado, K., Demkin, M., & Cukierski, W. (2019). *2019 Data Science Bowl*. Kaggle.

Feng, T. (2019). *Using game-based measures to assess children's scientific thinking about force.* [Poster session]. American Educational Research Association Conference, April 5–9, 2019, Toronto, Canada.

Ginsburg, H. P., & Baroody, A. J. (2003). *Test of early mathematics ability* (3rd ed.). ProEd.

Google Analytics. (2024a). *Analytics 360: PBS KIDS Games app* 20230701–20240630.

Google Analytics. (2024b). *Analytics 360:* pbskids.org (browser traffic only, excluding WebView browsers) 20230701–20240630.

King, A. (1991). Effects of training in strategic questioning on children's problem-solving performance. *Journal of Educational Psychology, 83*(3), 307–317. https://doi.org/10.1037/0022-0663.83.3.307

Lovato, S., Felline, C., & Roberts, J. (2023). *Measuring distance between solutions in an engineering game for children.* [Poster session] Society for Research in Child Development Conference, March 23–25, 2023, Salt Lake City, UT.

Mayer, R. E. (2023). Improving learning from screens for toddlers and preschoolers. *Journal of Applied Research in Memory and Cognition, 12*(4), 473–475. https://doi.org/10.1037/mac0000133

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/BM.2012.031

Nagle, R. J. (2007). Issues in preschool assessment. In B. A. Bracken & R. J. Nagle (Eds.), *Psychoeducational assessment of preschool children* (4th ed., pp. 29–48). Lawrence Erlbaum Associates Publishers.

Nanjappa, A., & Grant, M. M. (2003). Constructing on constructivism: The role of technology. *Electronic Journal for the Integration of Technology in Education, 2*(1), 38–56.

Nielsen NPOWER. (2024). L+7, 9/25/23 - 9/29/24, M-Su 6A-6A Reach (000), PBS stations, 50% unif., 1+ min.

PBS. (2024). *2024 PBS Trust Survey* [Flyer]. https://dc79r36mj3c9w.cloudfront.net/prod/filer_public/value-pbs-bento-live-pbs/Downloadables/935e535c5e_PBS%20Trust%20Survey%20Flyer_2024.pdf

Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Parks, C. B., Schenke, K., Michiuye, J. K., Choi, K., Ziyue, R., & Wu, Z. (2020). *Cat in the Hat Builds That analytics validation study* (Final deliverable to PBS KIDS). UCLA/CRESST.

Redman, E. J. K. H., Chung, G. K. W. K., Feng, T., Schenke, K., Parks, C. B., Michiuye, J. K., Chang, S. M., & Roberts, J. D. (2021). Adaptation evidence from a digital physics game. In H. F. O'Neil, E. L. Baker, R. S. Perez, & S. E. Watson (Eds.), *Using cognitive and affective metrics in educational simulations and games: Applications in school and workplace contexts* (pp. 55–81). Routledge.

Redman, E. J. K. H., Feng, T., Parks, C. B., Choi, K., & Chung, G. K. W. K. (2023). *Learning-related analytics KPI—KPI Final Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Redman, E. J. K. H., Parks, C. B., Michiuye, J. K., Suh, Y. S., Chung, G. K. W. K., Kim, J., & Griffin, N. (2021). *Social-emotional learning games validity study* (Exploratory study): Final study report. UCLA/CRESST.

Redman, E. J. K. H., Schenke, K., Chung, G. K. W. K., Parks, C. B., Michiuye, J. K., Feng, T., Chang, S. M., & Cai, L. (2019). *Analytics Validation Final Report* (Final deliverable to PBS KIDS). UCLA/CRESST.

Roberts, J. D., Chung, G. K. W. K., Feng, T., Riveroll, C., Redman, E. J. K. H., Schenke, K., Lund, A., & Rodriguez, J. (2019). *Deriving learning-related measures from game telemetry: Detecting children's alternative conceptions of the pan balance.* [Poster session]. Biennial Meeting of the Society for Research in Child Development, March 21–23, Baltimore, MD.

Roberts, J. D., Parks, C. B., Chung, G. K. W. K., Redman, E. J. K., Schenke, K., & Felline, C. (2018). Innovations in evidence and analysis: The PBS KIDS Learning Analytics Platform and the research it supports. In *Getting Ready to Learn* (pp. 231–248). Routledge.

Rodriguez, J., Arena, D., & Roberts, J. D. (2018). Adaptive and personalized educational games for young children: A case study. In *Getting Ready to Learn* (pp. 212–230). Routledge.

Rose, D. (2000). Universal design for learning. *Journal of Special Education Technology, 15*(4), 47–51. https://doi.org/10.1177/016264340001500108

Schenke, K., Redman, E. J. K., Chung, G. K. W. K., Chang, S. M., Feng, T., Parks, C. B., & Roberts, J. D. (2020). Does "Measure Up!" measure up? Evaluation of an iPad app to teach preschoolers measurement concepts. *Computers & Education, 146*, 103749. https://doi.org/10.1016/j.compedu.2019.103749

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes* (A. R. Luria, M. Lopez-Morillas, & M. Cole, Trans.). Harvard University Press.

Younger, J. W., Felline, C., Killian, R., Corrado, K., & Roberts, J. D., *Evaluating effectiveness of educational games in natural settings*. [Conference presentation abstract]. 2025 Digital Media and Developing Minds International Scientific Congress, July 13–16, 2025, Washington D.C., USA.

Younger, J. W., Roberts, J. D., Felline, C., Corrado, K., & Lovato, S. *The role of learning analytics at PBS KIDS*. [Poster session]. Biennial Meeting of the International Mind Brain and Education Society, July 10–12, 2024, Leuven, BE.

Credits

Privacy

PBS KIDS is committed to creating a safe and secure environment that family members of all ages can enjoy. Children's privacy and safety are our top priority. As such, PBS KIDS never collects personally identifiable information. Consistent with our privacy policy, we and our service providers (like Google Analytics) intend to only collect and analyze data that is needed to deliver the high quality educational media experiences that users expect and to operate necessary business functions. To view the full PBS KIDS privacy policy, please visit: pbskids.org/privacy/.