noggin

Growing Kids' Noggins

Implementing a Learning Impact Evidence Framework in a Multimedia Children's Platform

Kevin Miklasz, Michael H. Levine, Makeda Mays-Green, Courtney Wong Chin, Kiersten Zimmerman

Acknowledgement

We would like to thank everyone at Nickelodeon and Noggin who contributed to testing and refinement of this framework as we developed our content, including in loose chronological order: Winnie Cheung, Hannah Peikes, Dr. Israel Flores, Ilana Weiss, Evan Rushton, Rachel Kalkstein, Hannah Ally and Dr. Sofia Jimenez. We also want to thank those who helped advise on the framework: Dr. Elizabeth Owen and Dr. Daniel Hickey. We also want to thank Lasette Canady and Ji Cho for their work on the design of this paper.

This paper is adapted from a chapter in the *Next Generation Evidence* book published by Brookings Institution Press.

Published by Noggin© 2023. To contact us about this paper, email NogginResearch@nick.com

Suggested citation for this paper:
Miklasz, K, Levine, MH, Mays-Green, M, Wong Chin, CBV, and Zimmerman, K. 2023. Growing your Noggin: Implementing a Learning Impact Evidence Framework in a Multimedia Children's Platform. Published on www.noggin.com/research/.

Introduction

Driving evidence-based outcomes in early childhood education is an urgent national priority given the strong scientific research revealing the long-term value of preschool learning and the critical period of early brain development. National policy leaders are prioritizing evidence-based early learning program expansion as a part of their bipartisan agenda. Focusing on outcomes at large scale, however, is a relatively recent phenomenon. Its origins track back to the first National Education Goal for "readiness," which followed decades of debates about closing performance gaps, and many related waves of reform in the K-12 standards-based movement.

As a children's media organization, Noggin faces the challenge of developing evidence-based offerings that engage and delight our audience of preschoolers. These days, young children have a sea of choices in the digital kids landscape-- Roblox, Minecraft, Toca Boca, and Scratch are just the tip of the iceberg. To capture young minds, creators must be deft in blending fun and engagement with intentional, outcomes-oriented content. With the emergence over the decades of high quality educational media–Sesame Street, Mr. Rogers, Noggin, Nick Jr. and the Public Broadcasting Service–one silver lining in this "digital wild west" is parents' demand for educational brands that can help their children prepare for school and life. Additionally, families are emerging from several years languishing at home during the COVID crisis, adding urgency to concerns that children's media time be purposeful.

As we re-envision our work at Noggin–the early learning platform developed by Nickelodeon and Sesame Workshop two decades ago, and now a part of Paramount–research will play an increasingly key role in the content production pipeline. Research will help us determine if content resonates with and engages children, and whether it supports learning key concepts and skills. The latter research, known as "learning impact research," has a modest, but established tradition among scholars who study the potency of informal media, including professional journals devoted to the impact of the changing media landscape, landmark studies of Sesame Street's long-term impact on learning trajectories, and meta-analyses of the educational promise of long-form digital games.

The current state of learning impact evidence

Today, all learning products intended for children should have proven impact or evidence of encouraging learning. However, what counts as appropriate evidence is still evolving.
The implementation of the 2001 No Child Left Behind (NCLB) Act was the beginning of an increased focus on ensuring that educational technology content and products would produce learning.
The 2015 Every Student Succeeds Act (ESSA) replaced NCLB, tying federal funding explicitly to a set of standards for learning impact, known as the ESSA Evidence Tiers, a set of 4 levels of evidence that define what counts as rigorous. The level of rigor and quality of the evidence increases from Tier 4 (Demonstrates a Rationale) to Tier 1 (Strong Evidence).

The ESSA standards showed progress in thinking about learning impact evidence, but there has been some criticism, including lack of detail about how specific research meets each ESSA Tier. As a result, other agencies offer their own interpretations of how to translate the ESSA standards into practical guidance for researchers. But their interpretations are not in complete agreement (for example, see SIIA, WWC, Evidence for ESSA). A second and more significant criticism of the ESSA Tiers, is the lack of guidance for how they apply to protocols in development; currently the standards only apply to fully developed products. Platforms like Noggin continually release new content, but a point-in-time ESSA Tier 1 study can take two to three years to complete. This means, during the period of the study, an estimated 300-600 new pieces of content would be added to the Noggin, making the study results obsolete by the time it's complete.

As another angle, the U.S. Office of Educational Technology established protocols for how to use rigorous development practices that involve testing and iteration throughout development (e.g., The EdTech Developer's Guide, Expanding Evidence Approaches for Learning in a Digital World). The standards provide guidance on when and how to do this work and best practices, but do not advise on what counts as rigorous, nor offer any guidance on how to prove that a particular product's development process was rigorous.

Digital Promise offers Research-Based Certification, an approach to impact research that focuses on the organization's practices, and not on the product. The organization undergoes a process to certify that their development processes follow best practices found in the research literature. If successful, Digital Promise awards the organization with an open badge and gives acknowledgment on the Digital Promise website. The drawback to this approach is that the certification does not indicate whether an organization itself conducts good formative research on specific content. This would be much more challenging to certify at scale.

In an academic study, Hickey and Pellegrino (2005) describe three general approaches to thinking about assessment of learning impact. The first, an Empiricist approach, is about measuring facts and associations between them. The second is a Rationalist approach, which measures mental models that students build. The researchers note that large, long-timescale approaches have to use one of these two approaches, with the more traditionally rigorous the assessment, the more the assessment itself tends to rely on the Empiricist approach (the bread and butter of classic multiple choice achievement tests). Neither of the first two approaches is effective for in-the-moment measurements of learning. Thus, Hickey and Pellegrino offer a third Sociocultural approach, which is about seeing evidence of authentic dialogue and practice. It focuses on how a student uses their environment to engage with a knowledge or skill that increasingly mimics the way experts in that skill would also engage with their environment. Sociocultural assessments are more effective with shorter timescale and near-transfer assessments, which offer a particularly relevant model for rigor in formative research.

Bridging the gap: Using impact evidence in formative research with media

As a solution to the challenges of studying learning impact and rigor of content in development, the Noggin team has developed a framework that tests for learning impact throughout the lifecycle of a piece of content. This process enables us to identify learning evidence well before we have the time or resources set aside to run an intensive randomized control trial that produces Tier 1 ESSA evidence. Accordingly, we've developed three evidence levels described in Figure 1. Lower levels are considered less rigorous, but moving to one level lower is typically an order of magnitude less costly and time-intensive. Our general approach to impact research is to start by gathering lower levels of evidence, and once proven, spend time and resources investigating higher levels of evidence, thereby avoiding using large amounts of resources only to discover something doesn't work. Additionally, the lower level Directional Evidence research is effective with rapid cycle content iteration needs and ensures the content continues to improve as we develop it.

Name	Short definition	Criteria
Directional Evidence	Evidence is trending in the direction that impact exists	*Must show evidence that is consistent with the idea that learning growth is happening. The evidence is necessary but not sufficient.*
Correlational Evidence	Usage of the content is correlated with learning gain	*Must show learning growth is correlated to usage. That can be either through 1) showing that higher usage corresponds to more learning growth; or 2) that pre-post gains occur when using the content.*
Causal Evidence	Usage of the content causes learning gains	*Must show learning growth as a result of usage, as compared to a well-defined control group.*

Summary of the three Noggin Learning Impact levels. The following explores each level in more detail.

Directional Evidence

Directional Evidence indicates evidence that is directionally consistent with the concept that learning is happening. Directional Evidence can arise from 1) alignment between usage and best practices; 2) observations that learning is happening in the moment; 3) informal measurements that learning has happened over repeat play; 4) ability to transfer learning from the activity to a related task; or 5) a positive but insignificant Correlational or Causal Evidence. We choose one of these five approaches for Directional Evidence based on whatever makes the most sense given the nature of the content.

Directional Evidence is typically found in our formative research process or during the content development process on alpha or beta versions of content, but we can also look for this evidence post-launch. The techniques are light and quick forms of evidence-gathering that maintain elements of quality and rigor.

Directional Evidence is most similar to ESSA Tier 4 (Demonstrates a Rationale), but ESSA doesn't fully acknowledge in-process design research as a valid form of evidence. Our standard requires actual evidence, making it more rigorous than the ESSA Tier 4. Our approach is in the spirit and intent of this 4th level of ESSA, which is to acknowledge products that have not been directly measured for impact but there is good reason to believe they are effective.

Accordingly, our Directional Evidence is most strongly influenced by and derives its rigor from the Sociocultural approach advocated by Hickey and Pelligrino. All the levels of evidence typically involve identifying some form of authentic dialogue representing genuine engagement with the learning content.

Correlational Evidence

Correctional Evidence attempts to make a claim that usage is correlated with some kind of learning. There are two general categories that qualify for this level. The first directly proves a statistically significant correlation between a usage metric and a learning metric. Second is one where learning gains are seen from a pre-post measure, with usage of the learning tool interjected in between. This can be thought of as an intervention without a control group.

In either case the lack of a well-defined control group is the defining feature that results in correlation but not causation. "Well-defined" is the key phrase, and we mainly look to the ESSA standards for its definition. Correlational Evidence is most similar to ESSA's Promising Evidence. We generally follow the ESSA definitions with the exception that we do not require "statistical controls for selection bias" as that is an overly stringent requirement for a correlational study, arguably making ESSA's Tier 3 Evidence no different from ESSA's Tier 2 Evidence, as those statistical controls are what makes a control group "well-defined." We follow the SIIA interpretation of ESSA where the ESSA standards lack detail.

Causal Evidence

Causal Evidence strives to make a causal claim. The goal is to find that usage of a learning tool causes learning gains, typically in comparison to a control group. The classic form of this study is a randomized control trial, but many newer machine learning techniques are now considered to also make causal claims with various degrees of comparable rigor. One particular category of studies (often bundled as quasi-experimental studies) is one that defines control groups after-the-fact but does so in a way that ensures no selection bias in how the control group is defined and thus is a "well-defined" control group.

Our Causal Evidence category combines ESSA's Tier 2 and Tier 1 Evidence into one level, comprising quasi-experimental and "true" experimental (aka randomized control trial) approaches. The reason for the combination is because both are forms of causal studies and because several innovations in big-data-driven quasi-experiments (notably those using propensity score matching) are arguably more robust than limited sample size RCTs, making this distinction in methodology antiquated. We follow the SIIA interpretation of ESSA where the ESSA standards lack detail.

Practical application of the Impact Evidence Standards

Below is a brief description of Noggin's general content production process. We describe each step in general terms since each type of content on Noggin goes through a slightly different form of this process.

1 Background Research

The learning and content teams conduct background research on the skill, looking at best practices in the research literature.

2 Advisor Feedback

Upon determining what to produce, we engage our outside expert panel to review our ideas. Our robust advisory panel is composed of researchers and experts in early childhood education, representing professionals in academia and other media organizations.

3 Formative Research

During production, we conduct usability tests throughout the various stages of content development, typically at key "alpha" and "beta" stage milestones. The early stage tests may or may not test for impact evidence; during the late stage test we attempt to incorporate a Directional Evidence study.

4 Post-Launch Engagement Analytics

For the first few weeks after launch, we monitor basic engagement analytics. Although not testing for impact, this does indicate if the content is resonating or is unexpectedly unpopular, helping us identify issues to address.

5 Post-launch Learning Analyses

Several months after launch, we use child performance data to conduct a learning analytics analysis or do a deeper qualitative research test. This produces either Directional or Correlational evidence, depending on the format of the content and what data are available.

6 Summative Research Study

Considering the high investment required for summative research, we selectively employ summative research studies to test our content at large–either groups of sequential and related sets of content or the Noggin platform as a whole. This gives a broader view of our content that can produce Correlational or Causal Impact evidence.

Details about Noggin's Formative Research Approach

Noggin's approach to formative research reflects our approach to helping children grow and learn through applying the same principles as follows:

- **Putting Children First**

  Understanding the needs, preferences, behaviors, and goals of children and families to best support learning

- **Focusing on Impact**

  Delivering content that demonstrates learning objectives during in-development testing

- **Innovating**

  Exploring new research methods to develop a deeper understanding of children and families

Following are examples that demonstrate how we have been applying the above evidence standards to various types of content that Noggin has produced in the last few years.

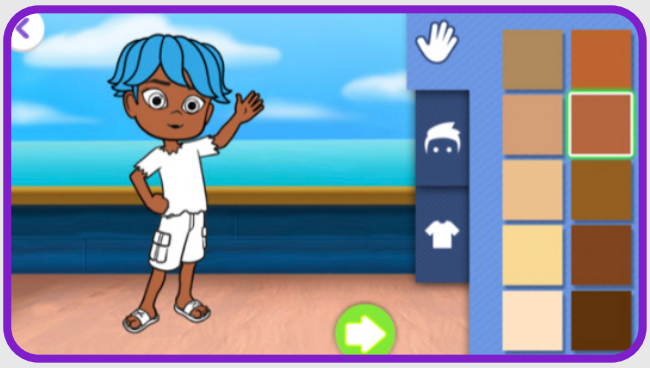
A Directional analysis of Yoga Friends series using formative research

A screenshot of the Yoga Friends series.

Yoga Friends is one of Noggin's classes aimed at supporting health and wellness. In this class, a yoga instructor guides children through a yoga flow (series of poses) in fun, animated settings such as a farm or the ocean. A learning goal of the series is that children would work on physical health by beginning to perform yoga poses, which involves practicing balance, strength, flexibility and breathing. Children watched one episode in session with researchers and two episodes at home with caregivers.

We found that all children attempted to do most of the yoga poses while viewing episodes and were able to demonstrate at least one of the featured yoga poses when prompted with the pose name or visual; thus, this series passed Directional Evidence. Following these and other findings regarding engagement, we aimed to strengthen the remaining episodes in the season by demonstrating challenging poses more than once and including a mix of breathing, balance (static) and active (movement) poses.

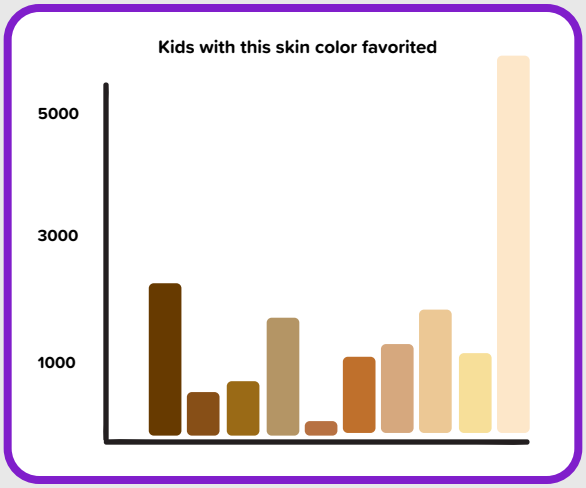
A Directional analysis of a Social and Emotional Learning game using learning analytics



A screenshot of the Friends: Create and Color Game

In the game Friends: *Create and Color*, a child chooses a property featured on Noggin (e.g., *Bubble Guppies*) and creates a character in the style of that property by choosing characteristics, such as hair style, hair color, skin color, clothing, eye type, etc. The learning goal of the game is that children will learn that friends can have different hair styles, skin colors, facial features, and preferences.

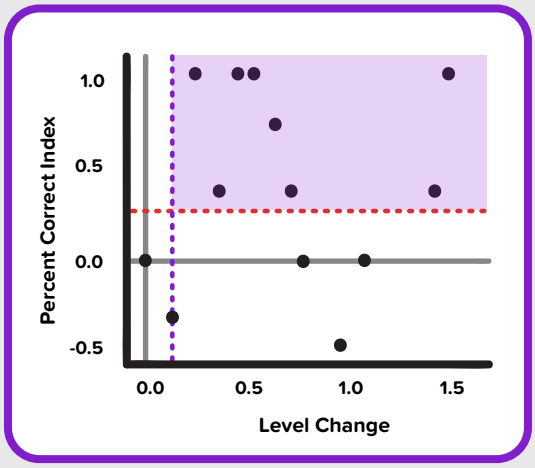
Through analysis of gameplay data, we found that the majority of children (70%+) made friends with different skin tones, hair styles and hair colors when replaying the game, so that the game passed Directional Evidence. But we also found that choices children made differed based on the property and interface, and that options selected reflect the population racial makeup. These results point to areas to adjust in the game to better target the learning goals.



Number of kids who showed a preference for different skin color choices.

A Directional analysis of Math and Literacy interactive content using learning analytics

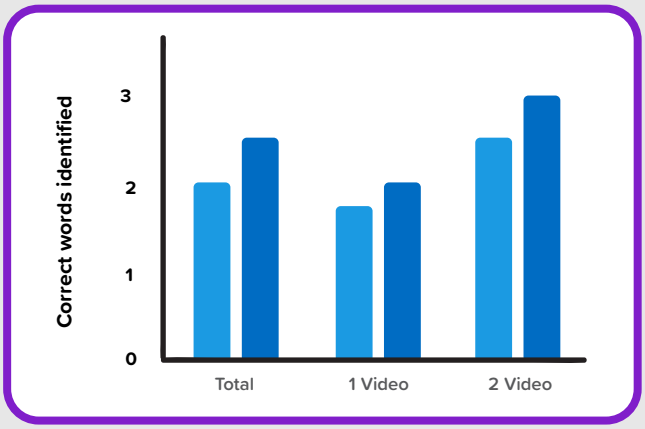
Noggins 123s and ABCs, a series featuring math and literacy content, begins with a short introductory video followed by an interactive activity that prompts children to practice the skill learned in the video by completing a series of right/wrong answer tasks. As children do well in the game, the games level up and deliver more challenging content on replay. Using gameplay data, we filtered for children who replayed these games, and then measured how much children leveled up on average, and for those that did not level up, how much did their performance increase on replay. From these two metrics we found that 62% of the games developed in 2022 had both statistically significant increased levels upon replay and also statistically significant performance increases for those that didn't level up upon replay. Thus this piece passed Directional Evidence. But we also found several pieces of content that didn't pass impact and fell into less positive regions. Those pieces went through additional learning analytics, and in some cases formative research review to determine why there was minimal learning growth. This led to recommended changes to those content pieces.



A graph showing where the first 13 Noggin ABC and 123 pieces of content fell on our two impact metrics. Pieces in the purple region showed a significant increase on both measures, and were deemed to pass Directional Impact.

A Correlational analysis of Noggin's Vocabulary series using summative research

Noggin has developed several different approaches to teaching vocabulary through short videos. To test the relative effectiveness of different video formats, we chose to study three vocabulary series: *Show Me Bot, Noggins at Work, and Word Play.* From each series, we chose three vocabulary words in order to expose children to nine words in total. We gave children a "Peabody Picture Vocabulary Test"-style pre- and post-assessment on the nine words.



Comparison of pre- and post-assessment test scores for Word Play. The right two comparisons identify kids that had one vs. two viewings of the content.

A screenshot from Word Play: Seaweed where a visual of the vocab is introduced on screen along with the written word present.

The study included 15 children who watched all of the videos one or two times (on consecutive days) before taking a post-test. Results indicated different effectiveness of the formats: *Word Play* had a positive and statistically significant effect; *Noggins at Work* had a positive but not quite statistically significant effect; and *Show Me Bot* had a neutral effect that was not statistically significant. Thus *Word Play* passed Correlational Evidence, and *Noggins at Work* passed Directional Evidence. These results led to an internal analysis of the creative differences between the video series, with reuse of elements from the two more successful series in later vocabulary videos.

A Correlational analysis of the content and leveling system in a Math game using summative research

A screenshot of the Tale of the Sleepy Knight game.

The math game Tale of a Sleepy Knight teaches matching and sorting. The game levels up offering increasing challenges as children do well and master matching and sorting skills. To test how the leveling-up system could enhance learning, we developed a pre- and post-assessment that tested the same skills in the game but on items that weren't used in the game. First, we had children take the pre-test; then they played the game at least once a week for three weeks; and finally they took the post-test.

We found that all children leveled up through the game but at different rates depending on their performance. Additionally, we discovered statistically significant increases on the post-assessment score after several weeks of play, with the largest gains among 3 and 4 year olds, indicating this game passed Correlational Evidence. This demonstrated that the game itself is effective at teaching the skills and that our basic replay and leveling system seemed to effectively enhance the learning.

	Number of Kids	Average Score Gain	*p* Value
Age 3	7	2.1	0.045
Age 4	10	2.8	0.001
Age 5	5	1.4	0.004

A chart showing the key pre-post test results broken down by age.

Conclusion

Best practices for developing learning content in the children's educational media industry have long been established by leaders, such as the Public Broadcasting System, Nickelodeon and Sesame Workshop. Additionally, research standards have also been established for understanding quality and rigor of developed and released content. This paper is one of the first attempts to join best practices and research standards to create a single framework that guides rigor in content research, from early development through content release and summative research. As children's media leaders, we believe that taking a proactive role in evaluating the learning impact of content throughout the entire development process is a necessary step toward ensuring the continued success and quality of children's content. We offer this model and these examples with the hope that others may find them useful or inspiring for their own work. We eagerly look forward to opening a dialogue about the best ways to use impact research to help children learn.