NOGGIN

LEARNING IMPACT EVIDENCE IN A MULTIMEDIA CHILDREN'S PLATFORM

KEVIN MIKLASZ, MAKEDA MAYS GREEN, AND MICHAEL H. LEVINE

Driving evidence-based outcomes in early childhood education is an urgent national priority: strong scientific evidence about the long-term value of preschool learning and the critical period of early brain development is now broadly understood. The new federal administration has made evidence-based, quality early learning program expansion a large part of its agenda.[1] However, a needed focus on outcomes is a relatively recent phenomenon, tracked back to the first National Education Goal for "readiness," which followed decades of debates about closing performance gaps and many related waves in the K-12 standards-based reform movement.[2]

As a popular media organization, Noggin faces significant challenges in developing evidence-based offerings that will not offend the tastes of our choosy audience of preschoolers!

These days, young children have a sea of choices in the digital kids landscape—from Pokémon to Minecraft to Toca Boca to Scratch—that engage their minds and bodies about three hours a day.[3] Creators must be deft in blending fun and engagement with intentional, outcomes oriented content. One silver lining in this digital wild west[4] is the demand from parents—going back several decades, with the emergence of Sesame Street,

Mr. Rogers, Noggin, Nick Jr., and the Public Broadcasting Service, for educational brands that can help children get ready for school and life.[5] And the present needs of young children, emerging from over a year languishing at home during the COVID crisis, have added urgency to concerns that media time be purposeful.

That is why, in retooling our work at Noggin, the early learning platform first developed by Nickelodeon and Sesame Workshop two decades ago and now a part of Paramount, research has become a key component of the content production pipeline. We use research not only to determine if content resonates with or engages children but also to learn if it helps them acquire key concepts and skills. The latter research, which we call "learning impact research," has a modest but established tradition among scholars who study the potency of informal media, including professional journals[6] devoted to the impact of the changing media landscape, landmark studies of Sesame Street's long-term impact on learning trajectories,[7] and meta-analyses of the educational promise of long-form digital games.[8]

THE CURRENT STATE OF LEARNING IMPACT EVIDENCE

At this time, it is well established that learning products used with children should have proven impact or evidence that those products incite learning. Yet, how we establish what counts as appropriate evidence is still evolving. Starting with the implementation of the No Child Left Behind Act (NCLB) first enacted in 2001, there was an increasing focus on ensuring that educational technology content and products would produce learning. The 2015 Every Student Succeeds Act (ESSA) took NCLB a step further by tying federal funding explicitly to a set of standards for learning impact. These are commonly known at the ESSA Evidence Tiers,[9] a set of four levels of evidence that define what counts as rigorous. As research moved from Tier 4 (Demonstrates a Rationale) toward Tier 1 (Strong Evidence), the level of rigor and quality of the evidence increases.

As much as the ESSA standards are a huge step forward in thinking about learning impact evidence, there have been (a few) criticisms of the standards. First, the government standards themselves were not written with enough detail to be clear on how specific research meets each tier. This has resulted in other agencies offering their own interpretations of how to translate the ESSA standards into practical guidance for researchers, and

their interpretations are not in complete agreement (for example, see SIIA,[10] WWC,[11] and Evidence for ESSA[12]).

Second, and most significant for our purposes, the ESSA tiers apply to fully developed products and there is no guidance for rigorous in-development protocols. For a product like Noggin that is a platform continually releasing new content, any point-in-time ESSA Tier 1 study would take one to three years to complete: an estimated 300 to 500 new pieces of content would be added to the Noggin platform during that time, and the study's results would become obsolete by the time they were finally released.

As another angle, the U.S. Office of Education Technology has laid out protocols for how to use rigorous development practices that involve testing and iteration throughout development (see The EdTech Developer's Guide[13] and Expanding Evidence Approaches for Learning in a Digital World[14]). Although the standards provide guidance on when and how to do such work and lay out best practices, there is no guidance on what counts as rigorous, nor any way to prove that a particular product's development process was rigorous.

Another approach to address this problem comes from the organization Digital Promise, with their research-based certification.[15] For this technique, the focus is less on the ultimate product and more on the organization. The organization undergoes a process to certify they have development processes that follow best practices found in the research; once the organization is successful, they are awarded an open badge and acknowledged on the Digital Promise website. The drawback to this approach is that the certification does not note whether an organization itself conducts good formative research on that content. This would be much harder to certify at scale.

In the academic world, Daniel Hickey and James Pellegrino have laid out three general approaches to thinking about assessment of learning impact.[16] They first describe an empiricist approach, which is about measuring facts and the associations between them, and, second, a rationalist approach, which is about measuring the mental models students build up. They note that large-scale, long-timescale approaches have to rely on one of these two models, and the more traditionally rigorous the approach, the more the assessment itself tends to rely on understanding facts and relationships between them (the bread-and-butter of classic multiple-choice achievement

FIGURE 4.9.1 A Summary of the Three Noggin Learning Impact Tiers

Tier	Name	Short definition	Criteria
Tier 1	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 probably not sufficient.*
Tier 2	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 better results, or 2) pre-post gains occur when the content is used.*
Tier 3	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.*

tests). For making in-the-moment measurements of learning, neither of these approaches is sufficient. Thus, Hickey and Pellegrino offer a third perspective, sociocultural, which is about seeing evidence of authentic dialogue and participation in a community of practice. Sociocultural assessments work better in shorter time scales and nearer-transfer assessments, which offers a particularly relevant model for rigor in formative research.

BRIDGING THE GAP: USING IMPACT EVIDENCE IN FORMATIVE RESEARCH WITH MEDIA

To solve those gaps, the Noggin team, which consists of an unusual blend of content developers, instructional design experts, and research scientists, has developed a framework that tests for learning impact throughout the life cycle of a piece of content. This allows us to find 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 have developed the following three evidence tiers, described in figure 5.8-1. Lower tiers are considered less rigorous, but moving down, one tier is typically an order of magnitude less costly and time-intensive. Our general approach to impact research is to start gathering lower tiers of evidence first and, once those are proven, spend the time and resources looking for higher tiers of evidence. This avoids having to spend large amounts of resources only to find out something does not work. Additionally, the lower tier research works well with rapid cycle content iteration needs, and ensures the content continues to improve as it is developed.

Let's go through each level individually.

Tier 1: Directional Evidence

This level of evidence indicates there is evidence that is directionally consistent with the idea that learning is happening. Directional evidence can come 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 typically is found in our formative research process or during the content development process on alpha or beta versions of content, but we also can look for this evidence post-launch. All the techniques are meant to be light and quick forms of evidence gathering that still have elements of quality and rigor to them.

In ESSA terms, directional evidence is most similar to ESSA Tier 4 (called Demonstrates a Rationale), but, really, ESSA does not fully acknowledge this kind of in-process design research as a valid form of evidence. The standard as we have written it is more rigorous than the ESSA Tier 4, as some form of actual evidence is required. This is in the spirit and intent of this fourth level of ESSA, which is to acknowledge products that have not directly measured impact but for which there is good reason to believe they are effective.

Accordingly, our Tier 1 evidence is most strongly influenced by the sociocultural approach advocated by Hickey and Pelligrino, and derives its rigor from that viewpoint. All the levels of evidence typically involve looking for some form of authentic dialogue that represents genuine engagement with the learning content being featured.

Tier 2: Correlational Evidence

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

In either of these cases, 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 the definition of this phrase. Correlational evidence is most similar to ESSA's Promising Evidence. We pretty much follow the ESSA definitions, with the exception that we do not require "statistical controls for selection bias," since that requirement feels overly stringent for a correlational study and, arguably, makes 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.

Tier 3: Causal Evidence

This evidence is attempting to make a causal claim. The goal is to say that the use 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) are ones that define control groups after the fact but do so in a way that ensures there is no selection bias in how the control group is defined, so it is, thus, a "well-defined" control group.

Our causal evidence category combines ESSA's Tier 2 and Tier 1 evidence into one level, which comprises both quasi-experimental and "true" experimental (aka randomized control trial) approaches. Both are combined 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 pipeline. We describe each of the steps in general terms, as each type of content we make goes through a slightly different form of this process.

Background Research

Our learning and content teams do background research on the topic, looking at best practices found in the field and research literature.

Adviser Feedback

After we have an idea of what we want to produce, we check our designs with an outside expert. We have a robust advisory panel, composed of researchers and experts in the early childhood education space, representing other professionals in academia and other media organizations.

Formative Research

Now we are in production. Usability tests are conducted throughout the various stages of content development, typically at key "alpha" and "beta" stage milestones. The early stage usability tests may or may not test for impact evidence, but at the late-stage test, we make every attempt to incorporate a Tier 1 impact study.

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, and may point to some issues to address.

Post-Launch Learning Analysis

Several months after launch, we will use the performance data to conduct a learning analytics analysis, or do a deeper qualitative research test. This can produce either Tier 1 or Tier 2 evidence of impact, depending on the format of the content and what data are available.

Summative Research Study

Considering the high investment needed for summative research, we selectively employ summative research studies to test our content at large, either groups of content that are meant to be sequenced and done together or our platform as a whole. This gives a zoomed-out view of our content that can produce Tier 2 or Tier 3 evidence.

As a practical example, we have mapped out the research life cycle for a recent piece of content: a vocabulary video series called Word Play (figures 5.8-2 and 5.8-3).

FIGURE 4.9.2 Word Play Research Life Cycle

1. **Background Research**
   The Word Play series was derived from our "Vocabulary" Skill in our Noggin Learning Framework, which intended to teach kids the meaning of specific words using simple engaging visuals and repetition.

2. **Advisor Feedback**
   Our advisors for vocabulary content include Dr. Susan Neuman of New York University and Dr. Glenda Revelle of the University of Arkansas.

3. **Formative Research**
   As a short form piece of video content lasting about 1 min in length, the production cycle was too rapid to do in-development testing. Instead we opted for a post launch testing described below.

4. **Launch Engagement Analytics**
   The Word Play had average launch statistics by both video starts and completion rates.

5. **Post-Launch Learning Analysis**
   Children were given a PPVT style vocabulary test as a pre-post measure and asked to watch a series of vocab videos at least once a day for two days. Scores showed a statistically significant increase from the pre-test to the post-test.

6. **Summative Research**
   We are planning to involve Word Play as one component in a larger intervention being planned now for later in the year, which aims to produce Tier 2 evidence for the entire set of content.

REFLECTIONS

The children's media field has had modest but notable success in designing content with measurable impact: industry best practices have been established by leaders such as the Public Broadcasting System, Nickelodeon, and Sesame Workshop. However, the earlier work was done prior to the emergence of learning standards and practices associated with evidence-based outcomes. As children's media leaders, we believe the next round of educational progress, in a post-COVID environment, will require a convergence in the expectations set by educators and content producers. It is our mission to help ensure this new approach is driven by digital teachers and role models that children truly love!

NOTES

1. Cody Uhing, "President Biden Seeks Important Funding Increases for Early Learning & Care Programs in FY2022," First Five Years Fund, April 9, 2021, www.ffyf.org/president-biden-seeks-important-funding-increases-for-early-learning-care-programs-in-fy2022/.
2. See The National Education Goals Panel website, https://govinfo.library.unt.edu/negp/page3.htm.
3. See Common Sense Media, https://www.commonsensemedia.org/sites/default/files/research/report/2020_zero_to_eight_census_final_web.pdf.
4. See Lisa Guernsey, Michael H. Levine, Cynthia Chiong, and Maggie Severns, "Pioneering Literacy in the Digital Wild West: Empowering Parents and Educators," Sesame Workshop, Joan Ganz Cooney Center, August 10, 2014, https://joanganzcooneycenter.org/publication/pioneering-literacy/.
5. See Our Impact, Sesame Workshop website, www.sesameworkshop.org/who-we-are/our-impact.
6. See Taylor & Francis Online, October 1, 2020, www.tandfonline.com/toc/rchm20/10/1.
7. Alia Wong, "The *Sesame Street* Effect," *The Atlantic*, June 2015, www.theatlantic.com/education/archive/2015/06/sesame-street-preschool-education/396056/.
8. See Douglas B. Clark, Emily E. Tanner-Smith, and Stephen S. Killingsworth, "Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis," *Review of Educational Research*, March 1, 2016, Sage Journals, https://journals.sagepub.com/doi/full/10.3102/0034654315582065.
9. See www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf.
10. Denis Newman, Andrew P. Jaciw, and Valeriy Lazarev, "Guidelines for Conducting and Reporting EdTech Impact Research in U.S. K-12 Schools,"

Empirical Education, April 15, 2018, www.empiricaleducation.com/pdfs/guidelines.pdf.

11. See "Using the WWC to Find ESSA Tiers of Evidence," IES WWC website, https://ies.ed.gov/ncee/wwc/essa.

12. See Evidence for ESSA, Find Evidence-Based PK-12 Programs, n.d., www.evidenceforessa.org/.

13. See "Ed Tech Developer's Guide," U.S. Department of Education, April 2015, https://tech.ed.gov/files/2015/04/Developer-Toolkit.pdf.

14. See https://tech.ed.gov/wp-content/uploads/2014/11/Expanding-Evidence.pdf.

15. See Certified Products page at the Digital Promise website, https://productcertifications.digitalpromise.org/certified-products-2/.

16. Daniel Hickey and James Pelligrino, "Theory, Level, and Function: Three Dimensions for Understanding Transfer and Student Assessment," Research Gate, January 2005, www.researchgate.net/publication/201381886_Theory_level_and_function_Three_dimensions_for_understanding_transfer_and_student_assessment.