Check for updates

A Framework for Research on Education With Technology

Daniel B. Wright*

Alder Graduate School of Education, Redwood City, CA, United States

Educational software offers the potential for greatly enhanced student learning. The current availability and political will for trying new approaches means that there is currently much interest in and expenditure on technology for education. After reviewing some of the relevant issues, a framework that builds upon Marr and Poggio's (1977) levels of explanation is presented. The research itself should draw upon existing cognitive, educational, and social research; much existing research is applicable. Guidelines for those conducting research and those wishing to acquire technology are presented.

Keywords: technology, CAI, EdTech, cognition, personalized learning, blended learning, cognitive tutor

While the phrases: blended learning, computer-assisted instruction, computer supported education, edutainment, e-learning, flipped classrooms, intelligent tutoring systems, interactive learning environments, personalized learning, serious games, teaching machines, etc., are relatively recent labels, people have been thinking about the cognitive processes upon which these phrases rely for millenia. Many of the issues now faced have been addressed before (e.g., how to provide teachers with clear actionable information about how their students are doing, creating enough content for students), though with modern computers some of the difficulties faced by, for example, the teaching machines of the 1950s, can be addressed more easily (e.g., allowing multiple response formats for questions).

The different labels that have been used over the decades and those currently used by different stakeholders convey subtle differences of focus and also reflect different marketing strategies. To avoid these a more generic phrase will be used here: *Education with Technology*, abbreviated EwT. This was chosen to stress that the emphasis is on education and that technology provides a method for implementing some aspects of education.

Suppes (1966) notes while Alexander the Great was able to have personalized tutoring from Aristotle, this privilege is not available to many. He argued that if the wisdom and skills of Aristotle could be delivered by a computer, this could be scaled to benefit many students. Training millions of people to become Aristotle-like personal tutors is not economically feasible. However, if computer software could be developed to perform like Aristotle for some tasks, the additional costs of scaling this up to allow many to benefit is relatively small if the hardware is in place and if the same program is suitable for many. Computer software is a very scalable technology. In the future there will be more technologies that can be used for education. Part of the success of any new system will be if it scales as well as computer software. For example, holodecks might be used in education (Thornburg, 2014), virtual reality glasses are already on the market, and neural implants designed to improve cognition are being built (e.g., https://www.neuralink.com/), but these would likely be expensive to scale-up. Nootropics (drugs designed to improve cognition) could become part of education discourse, and could be cost effective, but their use raises some ethical/health concerns. Future technologies could provide a radically different way to gain new information. For now, designing computer software is the most scalable technology available.

This first section is largely US focused. This is because the manuscript arose out of concerns for research and practice in the US. The framework is proposed for all EwT. The AIED (Artificial Intelligence in Education) conferences provide good snapshots of the relevant international research (André, 2017; Conati et al., 2015).

LEARNING BEFORE COMPUTERS

People have been interested in the cognitive processes that underlie learning and memory for centuries (e.g., Yates, 1966; Carruthers, 1990; Rubin, 1995; Small, 1997). Different theories of how people learn and remember have lead to different theories of how people should be taught (Roediger, 1980). For example, if you assume that memory works like a file cabinet, where memories exist unaltered, memory errors result from not finding the "right" file. It follows that educational approaches should attempt to put knowledge into these cabinets until they are filled and teach retrieval techniques. With this metaphor, memory errors of omission (forgetting) may be seen as failures of retrieval, but elaborate errors of commission (confabulation) should be rare. Other memory metaphors (e.g., a sponge, wax tablet, a paleontologist recreating a dinosaur) suggest different reasons for memory errors and different pedagogies.

Over the centuries memory researchers have examined both internal memory mechanisms and external memory devices, and how they interact. Rubin (1995) describes how internal memory aides, or *mnemonics*, were taught to those needing to recite long passages. An example is the method of loci where the person visualizes to be remembered information on a well learned path and then mentally travels along this path when needing to recite the information. These skills were viewed so important that their name is associated with a Titan in Greek mythology, *Mnemosyne*, the *mother of the muses* (the nine Muses, whose father was *Zeus*, were the sources of knowledge in the arts, sciences, and literature). In modern education there has been less emphasis on teaching students how to remember than on what to remember, though some recent textbooks include sections on how to learn the book's content (e.g., Nolen-Hoeksema et al., 2015).

External aides have also been used to facilitate learning and memory. For example, Yates (1966) describes several theaters that were designed to help enhance memory (e.g., Giulio Camillo's Memory Theater, the Globe Theater). These were designed so that the actors could mentally place what they later would need to recall in different parts of the theater. Many also believed that these theaters were designed to provide some additional magical value for memory. Usually technology should be considered an external aide, but it is important for students to be trained to use the technology. Distinguishing internal and external memory aides can be complicated. Clark (2008) argues that if a technology is always available and always relied upon that it is only biological prejudice that prevents someone saying that the technology is part of the person's mind. With technologies like Google Glass, neural implants, and nootropics, differentiating internal and external is complicated.

Language and writing are social (inter-personal) technologies that are important for education. Small (1997) describes how these were used in the creation of the first books. In oral traditions stories waned and flowed with the orator's and contemporary society's influence, but with books the story could remain unaltered for generations. People no longer had to rely on stories passed through many people as accurate representations of the original events. Human knowledge of *Atlantis* will have gone through many iterations before Plato wrote about it, but since then his writings have become record.

With the printing press, more people could read the same book. Most books are not personalized for each individual, but individuals with the economic means could choose which books they read. There have been attempts to personalize books and to introduce some control for the reader. Borges (1941/1998) describes this approach when critiquing *Herbert Quain's* fictitious novel *April March*. Borges (1941/1998, p. 109) used the schematic in **Figure 1** to show how a reader could navigate through *Quain's* novel. After a shared introduction (z) the reader chooses one of three y options, and for each of these the reader chooses one of three x options. An example of a complete story would be $z \rightarrow y_2 \rightarrow x_5$. This branching became popular in the 1970s with a genre of literature called *gamebooks*, where readers chose a path through the book by skipping pages. Two people could read the same book, but have different stories. *Quain's* novel, if it had existed, would have allowed readers to choose twice among three alternatives for nine possible paths. Borges (1941/1998, p. 110) said "gods and demiurges" could create systems with infinite paths. Near infinite branching is at the heart of many digital first person adventure games. Within education, so-called serious games also often use this branching to create different stories for different readers. One question is whether any positive aspects of allowing students the autonomy to choose their path outweigh any negative aspects of missing out on educational information from the paths that are missed.

Three other important technologies for education are radio, film, and television (e.g., Cuban, 1986; Ferster, 2014, 2016). These allowed what became known as *edutainment* to be heard and seen by millions. Initially there was much optimism. Thomas Edison declared in 1913 that "Books will soon be obsolete in our schools .... Our school system will be completely changed in 10 years" (as cited in Ferster, 2014, p. 32). Optimism is repeated
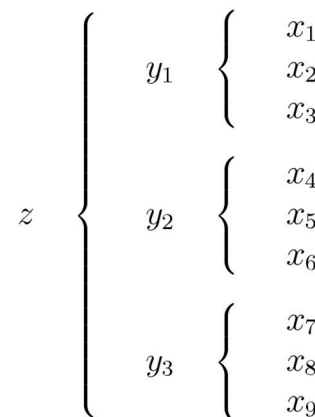

FIGURE 1 | Borges' schematic of *Quain's* novel *April March*. The reader can choose among nine different paths.

by some with the introduction of every new technology. There have been successes, but books are not obsolete. These media allow one production of a lecture to be provided to thousands of students, and the lecture is preserved for future students. These are mass media versions of the "sage-on-the-stage" approach to education (Ferster, 2014). Television meant users could simply switch on their edutainment. Over 500 million people did this for Carl Sagan's *Cosmos* in the 1980s. Shows for children, like *Sesame Street*, also have had large impacts.

The choice of content–when left to the whims of viewing statistics and advertisers' target markets–often will not lead to positive educational messages being broadcast. Sagan notes how society's choice for what to present via a variety of social media is unfortunate:

> An extraterrestrial being, newly arrived on Earth–scrutinizing what we mainly present to our children in television, radio, movies, newspaper, magazines, the comics, and many books– might easily conclude that we are intent on teaching them murder, rape, cruelty, superstition, credulity, and consumerism. We keep at it, and through constant repetition many of them finally get it. What kind of society could we create if, instead, we drummed into them science and a sense of hope?   (Sagan, 1996, p. 39)

He would find little solace with the content of the internet.

There are practical issues linking massed produced material onto formal courses. This can be done more easily when the material is for just a single course. In the United Kingdom, where the Open University has pioneered large-scale well-respected distance education since 1969, lectures were often on radio and television, and sometimes late at night. Nowadays students download their materials and this is also done with many massive open online courses (MOOCs). The Open University is a good example of an education system adapting their methods for distance learning with advances in technology. Home schooling has seen similar changes in relation to technology. The International Association for K–12 Online Learning (iNACOL, http://www.inacol.org/, "K–12" refers to the US grades kindergarten through 12th grade, which corresponds approximately to 5–17 years old), which began focused on home schooling, is now one of the main EdTech societies in the US. There are many large-scale courses available via the internet like Coursera, edX, Udemy, and Udacity, and stand-alone bits of knowledge that are available and used as part of educational courses including material from the Khan Academy, Wikipedia, and several YouTube (and YouTube-like) channels.

Another educational technology that pre-dates modern computers is teaching machines. In the 1920s and 1930s Pressey began creating machines to help to teach students. **Figure 2** shows a schematic of one of this teaching machines taken from a 1930 patent (submitted in 1928). Pressey presented these machines at American Psychological Association conferences and began selling them with the promise of "the freeing of teacher and pupil from educational drudgery and incompetence" (Pressey, 1933, p. 583). His teaching machines did not become popular. Benjamin (1988) and Skinner (1958) say part of the reason was the culture in the US at the time: "the world of
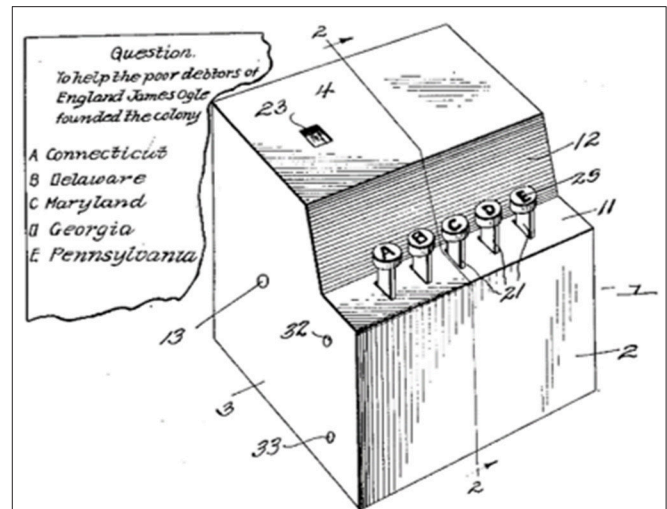

FIGURE 2 | From a 1930 patent by Pressey. The answer is D. James Ogle(thorpe) founded the colony in Georgia. From US patent "Machine for intelligence tests" (US 1749226 A).

education was not ready for them" (Skinner, 1958, p. 969). At the time Pressey was marketing these machines there was a surplus of teachers so there was less need for time-saving technology. Pressey blamed lack of sales on the overall economic depression (Ferster, 2014, p. 60).

When Skinner re-introduced teaching machines with some modifications in the late 1950s and US culture was more receptive. More teachers were needed to cope with the baby-boom and the US had just seen the Soviets launch *Sputnik*. There was a realization that the US needed to catch up with other countries particularly with respect to science education. President Eisenhower coined the phrase the *Sputnik Crisis* to describe this.

Skinner's (1958) approach differed from Pressey's, though they both assumed student-centered learning. Some of the differences were due to advances in learning theory within behaviorism as well as other areas of psychological research (e.g., Vygotsky's, 1978 work suggests a step-by-step approach through each individual student's zone of proximal development), and part Skinner's own nuanced approach. While Pressey was careful to say his machines would be a tool to help the teacher, Skinner was more comfortable saying his machines could do tasks formerly reserved for teachers: "the effect upon each student is surprisingly like that of a private tutor" (Skinner, 1958, p. 971). His claim lead to the popular press suggesting that his machines could lead to robots teaching students in classrooms like the research assistants taught pigeons in Skinner's lab.

Despite initial commercial success for some of the teaching machines of this era, their popularity faded. Benjamin (1988), Ferster (2014), and others discuss many of the factors that negatively affected their popularity. Three stand out: fear of technology, costs, and effectiveness. The first is that some people worried about students being taught by machines. While the *Sputnik Crisis* lead some to embrace technology, many feared that machines could create a dystopian future. These teaching

machines were being marketed at around the time that Ginsberg (1955/2014, p. 17) tapped into this fear referencing the Canaanite god of child sacrifice in his classic poem *Howl*: "Moloch whose mind is pure machinery!" Ginsberg was describing the dark-side of over-industrialization. The second reason is economic. These machines were expensive and more important for today's arguments, building each new machine was expensive so these did not scale-up as well as today's software solutions. Third, much of the research was showing that these machines were not as effective as initially promised.

EDUCATION WITH COMPUTERS

Papert (1993) began *The Children's Machine* by asking readers to imagine two groups of time travelers from the nineteenth century. The first are surgeons who are shown a modern surgery. Almost everything will appear new. The other group of time travelers are teachers who are shown classrooms of students sitting in rows listening to a teacher. While they would notice some differences in the classrooms, much would appear familiar. Ferster (2014, p. 1) repeated this thought-experiment two decades later: "a nineteenth-century visitor would feel quite at home in a modern classroom, even at our most elite institutions of higher learning." Why would modern classrooms seem so familiar to the nineteenth century guests? Is it that education got it right back then and that further advances were not necessary? Papert argued in letter to President Carter that education can be radically different and better if technology is embraced. He said: "Unless we do this, tomorrow will continue to be the prisoner of the primitivity of yesterday" (Papert, 1980).

There are several ways to classify different types of interactions students can have with an educational computer system. Atkinson (1968) and Suppes (1966) describe three: drill and practice, tutorial systems, and dialogue systems. Drill and practice can be seen as computer extensions of most of the early teaching machines. Students could take, at their leisure, practice quizzes and be provided with immediate feedback. Given the value of practice, testing, and feedback (e.g., Roediger et al., 2010), that different students will be best served by items that vary in difficulty and pertain to different competencies (Metcalfe, 2002), and that this is a monotonous task for teachers to do, drill and practice is an obvious part of the curriculum for computers to assist. Drill and practice systems allow students to evaluate their own knowledge efficiently.

The goal of tutorial systems goes beyond just allowing students to evaluate their knowledge. The goal is to teach students how to solve problems. The computer system can offer more interaction and feedback than a textbook or sage-on-the-stage edutainment, and more individualization than a teacher in front of a large class. For example, if most of a class has mastered calculating the area of a triangle, the remaining students could use a tutorial system to provide them with an alternative mode of teaching while the rest of the class learn a new task (which the students re-learning about triangles may or may not eventually be taught). While the teaching machines of the 1950s and 1960s could guide students, step-by-step, through different exercises, the computer

allows many more steps and allows the student to progress down multiple pathways (Ritter et al., 2016).

Dialogue systems allow a greater amount of interaction between the student and the system. Suppes (1966, p. 219) gives the example of a student asking: "Why are demand curves always convex with respect to the origin?" Fifty years on there have been many advances in natural language processing. I entered this phrase into Google and it suggested several web pages, including `quora.com`, where the question:

> What are the conditions under which a demand curve is convex? Explain with a few real life examples of goods with convex demand curves.

was asked and answered. Dialogue systems require some natural language processing. AutoTutor by Graesser and colleagues (for a review see Nye et al., 2014) is an excellent example of using language processing.

Papert (1980) describes another way to differentiate technology uses in education: auxiliary and fundamental computer uses. Auxiliary uses are where the computer is not changing the educational processes. The same (or very similar) activities are simply being presented in a different medium. These can be helpful, perhaps allowing individuals to work on their own activities and at their own pace, or making feedback more rapid. Using computers changes *how* the lesson is taught and the physical implementation of it, but there is not a major pedagogical shift. Fundamental uses change *what* is being taught and *why*. They enable students to learn information that they might not learn in a traditional classroom and learning is done in a manner that is a departure from the traditional pedagogy. Fundamental uses change the curriculum rather than implementing the same curriculum differently. While auxiliary uses can be beneficial and efficient, Papert argues that the fundamental uses have the potential to revolutionize education. In an essay written with former West Virginia governor Gaston Caperton, Papert describes how technology should be used not just to solve problems of "schools-as-they-are," but to build schools into "schools-as-they-can-be" (Papert and Caperton, 1999). This idea is at the core of the XQ Super School Project (https://xqsuperschool.org/), who fund proposals to create innovative schools.

Because computers are part of modern society, they will remain in students' homes and classrooms. However, the progress for educational software is neither straight-forward nor without impediments. A common comparison is made with the wild west (e.g., Reingold, 2015). In the wild west people could make wild claims about the curing powers of anything (e.g., heroin was marketed as a cure for cancer, sluggishness, colds, tuberculosis, etc., www.narconon.org/drug-information/heroin-history.html, Accessed June 21, 2017). If a vendor sold a lot of the product, a lot of money could be made, and in that era some people viewed making a lot of money as an important indicator of success. It was difficult for the public then to verify any of the these claims. A customer might be choosing a remedy for a sick child based on hope and desperation. Cuban's (2001) book title, *Oversold & Underused*, summarizes the view of many about

the impact of EwT. Schools want a computer system that will cover their entire curriculum and for all grades, and to improve scores on standardized tests immediately. They hope that there is just a switch to flip much like those in the wild west hoped a sip of a magical elixir would be a cure-all for any ailment. They hear a sales pitch that seems to offer this. It is important that those making decisions about EwT do not feel like they are making decisions out of desperation like the parents of a sick child in the wild west, but there is pressure on school administrators to have their schools move up in the rankings (Foley and Goldstein, 2012; Muller, 2018) and offering hope without evidence is a popular and persuasive sales technique in unregulated markets.

THE FUTURE

Technology is advancing. The time traveling teachers from Papert's and Ferster's examples would be amazed to see the number of students with cell phones, the capabilities of these devices, the amounts the devices get used, the ubiquity of social media (and its impact), and in general the technology related behavior of these *Digital Natives*. Some aspects of these affect EwT. For example, the small screens put constraints on the amount of text that can be shown at any one time just as writing on paper vs. animal skin affected what was written (Small, 1997). Detailed plots, long tables, and lengthy well-constructed arguments have been replaced by *tweets*.

Greenfield (2015) describes the phenomenon of *Mind Change*. The mind and the brain are adaptive to their environment. Many aspects of using computers (e.g., rapidly accessing lots of information in small pieces, social networking, "likes" on Facebook) make different demands on humans than traditional environments. She argues that it is important to research possible changes–some may be positive and some may negative–on the brains/minds of *Digital Natives* caused by new technologies. The EwT industry is betting that the positive effects greatly outweigh the negative effects, but consider the EwT approach of "asking Google." Does the rapid access to (possibly accurate) related information change the way people create questions and evaluate answers? Does the impersonal way that people get feedback from electronic tutors affect how the graduates would handle workplace criticism? Does the anonymity of the internet affect us as social animals? EwT can be implemented in different ways and each of these may affect students in different ways.

In order to predict future use of EwT accurately and to develop EwT well it is necessary to understand how EwT and education in general are situated within political, social, and economic climates. Convincing people to change their behaviors can be difficult. Pressey argued that the poor economic period in which he was creating his teaching machines meant that they were not financially successful. In the late 1950s and 1960s, when Skinner and others were creating teaching machines, the economic situation was better. Further, in the aftermath of *Sputnik* there was a societal drive to increase education. Still, the teaching machines of this time failed to have a lasting impact.

Currently, while the US is in fairly good financial shape, there is uncertainty about Federal funding of education research. Further, the decision making about buying specific products is non-centralized. This means that an attractive sales pitch is critical to the product's success. The start-up mentality is also evident. Many products are being developed with the backing of venture capitalists who hope that they have bet on some successful ones. Success of a single product that a venture capitalist bets on usually is greater in financial terms than the amount lost by several failed bets. Whether this gambling ratio is appropriate for education is debatable. In this climate Papert's (1980) notion of having a few centers of research is welcomed, though how they are funded and if they can maintain independence are uncertain.

The political climate in the US and elsewhere is divided with respect to scientific evidence based decision making and argumentation. While people benefit from science (e.g., the popularity of cell phones), many people are not interested in the science itself. If the magazines at supermarket checkouts are any indication, then the public has more interest in where reality stars vacation than scientific progress. The division between decision making based on science and based on superstition has long existed (MacKay, 1841/2012). Advanced technology has the potential for positive change, but it is necessary to make sure that the people embrace science over mysticism.

> … people use electricity and still believe in the magic power of signs and exorcisms. The Pope of Rome broadcasts over the radio about the miraculous transformation of water into wine. Movie stars go to mediums. Aviators who pilot miraculous mechanisms created by man's genius wear amulets on their sweaters. What inexhaustible reserves they possess of darkness, ignorance, and savagery! (Trotsky, 1933, October, 1933)

Trotsky was talking about how the scientific conditions among much of the population in Germany helped Hitler come to power, but parallels can be made with other places and time periods when a sizable proportion of the population lacks trust in the scientific method and when leaders who do not use valid evidence for their decision making come to power. These are not good circumstances to use scientific results to convince many in the public of the value of using technology in education.

Importantly, some people do value technology and do believe in its potential. There will continue to be investment on EwT. It is important for the research and the education communities to help target this investment. There are lots of products, many without much evidence of effectiveness. To prevent *survival of the loudest* dictating the evolution of EwT, it is important to think carefully about a potential EwT research framework. In the interest of children's education decisions should be based on the available science, rather than on, using Trotsky's phrase, "darkness, ignorance, and savagery!"

A RESEARCH FRAMEWORK

Billions of dollars are spent each year on technology for education. However, the current landscape is problematic. It

is important to go beyond the wild west metaphor (Reingold, 2015). This section is divided into three parts. The first section describes an attempt to discover whether a particular type of EwT is effective. The conclusion from this section is that identifying effective products is difficult. While many view a randomized controlled trial (RCT) of a product as the gold standard for evaluation research, here it is argued that there is a time and a place for RCTs, but that other research methods should also be used. The second section describes different levels of explanation. It is argued that EwT research should focus on the goals of the system and whether the underlying rules used to build the system effectively achieve these goals. These levels are based on an influential neuroscience framework put forward by Marr and Poggio (1977). The third section provides more detail about how research could progress, and provides some examples.

Testing the Effectiveness of Any EwT System Is Difficult

The US Institute of Education Sciences (IES) rightly states that "well-designed and implemented randomized controlled trials are considered the "gold standard" for evaluating an intervention's effectiveness" (https://ies.ed.gov/ncee/pubs/evidence_based/randomized.asp, Accessed June 22, 2017). However, it is often difficult to conduct an RCT of a product or any complex educational innovation in development. Many studies described as RCTs by their authors have significant problems and probably should not be called RCTs (Ginsburg and Smith, 2016). Sullivan (2011) discusses how forcing a research question into an RCT can distance the study from the intended experiences of the product/system. The argument here is not to avoid full-program evaluations. These can be very important in providing evidence for the effectiveness of well-established products. A good example is Pane et al. (2014) study of Carnegie Learning (Ritter et al., 2016). Randomization is useful, but is neither necessary nor sufficient for making causal inference (Wright, 2006; Pearl, 2009; Deaton and Cartwright, in press).

Performing experiments on components of the product can often be done more easily than evaluating whether and entire program works or not, and this approach can be beneficial for product development and may generalize to other products. Quasi-experiments still have their place, particularly with archival data. In the remainder of this section a study with the goal of evaluating the effectiveness of personalized learning (PL) is discussed to illustrate the difficulties of program evaluation.

The phrase "personalized learning" is often used to describe a wide variety of approaches (Arney, 2015; Horn and Staker, 2015; Taylor and Gebre, 2016). The core elements are that individual students decide some of the content and pace of their own learning, and that the system (usually a computer) guides and may restrict choices. This has the important consequence of freeing up time and resources so that the teacher can work one-on-one with each student or with small groups of students when they are not working on computers.

The study has achieved much press and optimism from investors (e.g., www.chalkbeat.org/posts/us/2017/05/22/as-ed-reformers-urge-a-big-bet-on-personalized-learning-research-points-to-potential-rewards-and-risks/. Accessed June 22, 2017). Pane et al. (2015) were funded to evaluate whether, in a nutshell, PL works. I describe several hypothetical "what ifs" and conclude that even if they had performed an RCT the results would have been difficult to interpret.

The authors highlighted how difficult a task it is to design a study to measure the effectiveness of PL and cautioned others not to over-interpret their results. Here is what they did. The schools in their PL condition had about 2 years of PL. These schools are described as those which "embrace personalized learning," "have a high degree of integrated technology as part of their school designs," are among "the country's best public charter schools," and have gone "through a series of competitive selection processes" (Pane et al., 2015, p. 36). These descriptions make these schools sound great! From these descriptions it might be expected that if a random selection of students were sent to these schools overall this group would perform better (i.e., raise their test scores by more) than if these students had been sent to a random selection of other schools. Rather than choosing a comparison group of schools with similar positive characteristics, Pane et al. (2015) had the test vendor (NWEA) match each student in these select schools with students at a variety of schools that presumably, on the whole, do not have all the positive characteristics described above for the PL group.

Given that interest is in school effects on student outcomes the decision not to compare similar schools is problematic. Even without an intervention the expectation would be that the PL group's scores should increase more because according to the authors' descriptions these are better schools than most. A good analogy would be if you were comparing restaurants. In one condition you have restaurants that are "the country's best" and in the control condition you have a random sample of restaurants (the schools were matched on urban, suburban, and rural, so restaurants might be matched on serving French, Italian, or Spanish cuisines). You gave each restaurant the same set of ingredients. For the "best" group you also gave them a recently published cookbook. The restaurants prepare meals using only the ingredients that they were given. Judges grade these meals and the "best" restaurants get higher marks. The question is whether you would conclude the cookbook was the cause?

If the PL group of schools were shown to have a positive effect in a study with properly matched group of schools (or if schools were randomly allocated either to have PL for 2 years or to be in a control group), what would this tell us and what would be the next steps? The norm in science when trying to establish causation is to have the treatment nearly identical for all units. In this study, however, "innovation was encouraged" for the PL schools. The schools were "not adopting a single standardized model of personalized learning" (Pane et al., 2015, p. 3). While this may or may not be beneficial for the education of the students in these schools, it makes it difficult for the researchers. Because of the variation in what PL means to people and how it was implemented among the PL schools, coupled with variation in teaching philosophies among the set of control schools, it would be difficult to conclude anything other than these hodge podges of difficult to characterize pedagogies may differ. While the

"positive" outcome on performance reported in Pane et al. (2015) would be predicted just because of how schools were sampled (there are also issues with respect to how students are allocated to schools), an RCT of a complex intervention would be unlikely to shed much light on *why* the intervention works. If an RCT (or a well designed matched-group study) showed substantial positive results, the next step would be to try to understand which components of the curriculum may be effective and research these components. An alternative is to begin with this research while these approaches are still being developed.

The purpose of the preceding paragraphs was not to criticize Pane et al. (2015). They did well in their attempt to answer a difficult question: "Does the set of things called PL work in the schools that are funded to do it?" Consider a simpler research question:

> What is the evidence that an EwT approach should work?

The remainder of this paper will explain what this question means in more detail and will argue for why it is a useful question for those seeking to purchase EwT and why it is a useful approach for developing research.

Levels of Explanation

Marr (1982/2010) reviewed vision research from the 1960s and 1970s. He marveled at research examining the physiology of vision, for example neuron firing patterns, but felt this did not provide a complete understanding of vision. To understand a system as complex as vision he argued that it was necessary to understand the system at multiple levels. He said that it was necessary to understand the goals of the system and the rules that the system used to achieve these goals in order to understand the system. Poggio (in the *Afterword* of Marr, 1982/2010 and in Poggio, 2012) discusses some changes that he recommends to the levels that he and Marr and had originally proposed (Marr and Poggio, 1977). He discusses how any classification system is somewhat arbitrary and that alternative levels and labels may be more appropriate in other contexts. He also stressed the importance of understanding the relationships among levels. This is particularly important for developing EwT. Here are five levels proposed for understanding EwT that are adapted from Marr and Poggio's (1977) framework.

1. Decide the top-level GOALS of the system. These will often be related to student learning, but may also be skills acquisition, behavior modification, etc.
2. Decide WHAT is to be learned. This might be something specific like the physics of volcanoes or a specific component of socio-emotional learning, or it may be broad like all academic subjects included in state tests or improving maladaptive behavior. Once these top two levels are decided researchers and developers can concentrate on how to achieve the goals for what is to be learned.
3. The CORE features of the theory for how the goals can be achieved. These will likely be from pedagogical or learning science theories. These features would include a "soft core" that can be evaluated, refuted, and adapted. Researchers

should continue to question whether these features produce the stated goals at the top level.
4. Decide the RULES or algorithm that will be used to set up the conditions that the core features of the theory predict will achieve the goals. These rules should be written in enough detail to allow them to be programmed into a computer language, to allow a carpenter to build a mechanical teaching machine, etc.
5. Decide the physical IMPLEMENTATION. For computer technology this would include choosing among tablets, smartphones, and desktops. The physical implementation will dictate, to some extent, how the rules are represented. With computers, this usually means the computer language used.

Once it is decided, for example, that the goal is for students to learn the physics of volcanoes, the researchers and developers would list some core features of the theory that they believe account for learning. Specific rules are developed that set up the conditions that these core features predict should increase learning. These are translated into a representation compatible with the physical implementation. This would be a top-down way to develop a product. Bottom-up development can also occur. Because of the widespread availability of computers and the scalability of software, some investors may only invest in computer software technology. The developers might then decide what they can build with this technology that may be profitable. The choice, for example, between building a product for foreign language learning vs. for learning physics may be based on what can be built with the physical device (e.g., foreign language tutorials usually require audio input/output and physics tutorials can be helped by allowing the user to manipulate diagrams).

Table 1 shows how the Pressey's and Skinner's teaching machines might fit within the proposed levels. At the top two levels the goal of the proposed technology is to increase student learning of the meanings of biology terms (other "whats" could be used). The teaching machines were built assuming the core features from the learning theories of behaviorism, developed by many of the psychologists of the time (e.g., Hull, Pavlov, Skinner, Thorndike, Watson). A rule that would increase the associations between biology terms and their definitions would be to present

TABLE 1 | Levels for Teaching Machines.

Levels	Description		
Goals	Student learning		
What	e.g., Biology terminology		
Core	Behaviorism (reinforcement, learning theory, associations)		
Rules	e.g., drill and practice		
Implementation	Teaching machines	Flashcards	Computers Tablets Smartphones

The top two levels (Goals and What) are assumed by the researchers and developers. The bottom three levels (Core, Rules, and Implementation) are designed to achieve these goals. Research should be done to see how processes at each level support higher levels.

the definition to the students and have them respond with the term, until they are correct on each item. This is a "drill and practice" procedure as described before. How to represent this rule will depend on how it is physically implemented. Suppose that the accuracy of a student's response does not require perfect spelling of a term; it would be necessary to have that described in the rules. With teaching machines the students might be presented with the correct answer and have to judge whether they were correct. With computers it is possible to use approximate matching algorithms to allow for mis-spellings, but the designer would still need to state how close a spelling could be (for example, the Levenshtein distance [LD]). Partial credit could be given for some answers, and details of this rubric would be required. The choice of physical implementation could also be constrained by higher levels. For example, if it were necessary to show a video or play audio, then flashcards could not be used.

The rules should be written out in detail. Consider a simple algorithm:

Let there be a set of items to be learned.

Let R_j be the number of times the student "correctly" answers the jth item.

Loop until $R_j = 2$ for all j.

    Sample one item from set of items such that $R_j < 2$ and ask student.

    If the student is correct add one to R_j.

Move onto next task.

These rules are simple enough that they could be implemented with different technologies, some of which are listed in the bottom row of **Table 1**. The different implementations should all use the same rules as specified, but could differ on aspects that are not specified. For example, how the sampling of items with $R_j < 2$ is done would vary depending on the implementation. Constructing a mechanical teaching machine for this purpose could be done, and the sampling would depend on how the gears work. A student using flashcards might have separate piles for the number of times the item was correctly answered (a pile for $R_j = 0$, a pile for $R_j = 1$, and a pile for $R_j = 2$) and the student might progress through the $R_j = 0$ then the $R_j = 1$ piles in the order the cards are in, placing a card on the bottom of the appropriate pile after its use. Computer software might use a pseudo-random process to shuffle all $R_j < 2$ items or create an order to optimize learning by spreading out semantically related items. If the sampling method turns out to be important for achieving the computational goals, then it should be specified by the rules, and this could constrain the choice of physical implementation.

These implementations can be judged on how well they implement the rules and the system's cost. Given the simplicity of this example all of these technologies should implement the rules accurately, though gears in teaching machines can break, flashcards get wet, and computers crash. Some implementations may have additional benefits built in, like making flashcards requires the student to write each item or that the computer can display information in innovative ways. The cost varies considerably among the different implementations. Creating

mechanical teaching machines for each student would be expensive and would be limited in what else they could do. The flashcards would be cheap to produce. They are often produced by the students themselves. Assuming the student already has a computer, making the software for a good "drill and practice" task can be expensive for the original prototype, but making copies available for additional users can be done inexpensively.

It is worth stressing that there will be important issues constructing the input and output for computers vs. tablets vs. smartphones. Research to make sure these modes are compatible is necessary, but these likely could implement the underlying rules in a similar way, perhaps changing how information is displayed. Further, this type of research (called human computer interaction, human centered technology, or user design) is well known by software companies so most EdTech companies have people in place for this. Compatibility research to examine, for example, if there is an advantage taking the SAT on a computer or a tablet, is already done by testing organizations.

Papert's auxiliary-fundamental (or "school-as-it-is" vs. "school-as-it-can-be") distinction is also interesting from the perspective of these levels. The auxilary uses would have the same upper levels as their traditional classroom counterparts. They might also assume the same core features of the learning theory and may even use the same rules. The physical implementation would differ from the traditional teaching curriculum. Auxiliary uses can still be valuable as the new technology may mean students learn more efficiently (e.g., by rapid feedback, having more one-on-one time with the teacher), but WHAT they learn would be the same. An example would be taking a well-constructed textbook, an item bank for each chapter, videos of excellent lectures, and changing the physical implementation of these so that they are delivered on a computer screen and through headphones. For fundamental uses WHAT students learn is different from the traditional curriculum. They may even introduce new GOALS.

As discussed earlier, PL is currently receiving a lot of attention. While there is not a single PL approach, **Table 2** shows a generic PL approach. Often PL is designed to improve student learning, broadly defined. The WHAT is often most of the academic

---

TABLE 2 | Levels for Personalized Learning.

Levels	Description		
Goals	Student learning		
What	All academic subjects		
Core	Individualization, choice, competency		
Rules	Allow the student control over the pace and content.		
	Estimate achievement and provide feedback to student.		
	Sometimes restrict student choice		
Implementation	Computers	Mentors (e.g., student to student)	Private tutor

The top two levels (Goals and What) assumed by the researchers and developers. The bottom three levels (Core, Rules, and Implementation) are designed to achieve these goals. Research should be done to see how processes at each level support higher levels.

curriculum. This was true for the sample of PL schools in Pane et al. (2015). **Table 2** shows three of the core features for why PL is assumed to increase student learning. The first feature, individualization, is that students are taught what is optimal for them. What is best for one person is not best for everyone. For example, the information should not be too easy or too difficult; it should be within what is called the zone of proximal development (Vygotsky, 1978) or the region of proximal learning (Metcalfe, 2002). The material might also be individualized to the students' interests and learning styles. The second core feature is that allowing students choice in itself is important for their growth. However, because students do not generally choose the items that will maximize learning (e.g., Metcalfe, 2002), this feature can conflict with the first feature. The third feature is what is often called competency based education, where students progress based on their performance rather than their age.

The rules for achieving these core features (which in turn are assumed to achieve the over-arching goal) with PL are often: allowing the student some control over the pace and content of their curriculum and using feedback to help them make good choices. A more detailed set of rules would be necessary to describe any specific approach. Suppose the system allowed students to choose a module and once within a module the students could decide (with feedback from the computer) whether they knew the module well enough to move onto the next module. Students would use the computer to try to learn the tasks, receive feedback on their progress, and then decide if they think they know the information well enough to move to the next task.

```
Loop
  Study information (amount and method determined by student)
  Take assessment
  Computer estimates student achievement
  Receive feedback
    If achievement estimate less than proficient, do not allow student to leave loop.
    If achievement estimate at least proficient, allow student to leave loop.
      Student decides whether to leave loop or repeat. Move to next task.
```

It can be beneficial to draw causal diagrams, which are called *directed graphs* in the branch of mathematics called graph theory, as in **Figure 3**. This is just a single module. It would be nested with many others into a course. This module may have pre-requisites or be a pre-requisite for other modules. It is important to consider how a localized causal diagram like **Figure 3** fits within a larger causal network (for more details about using graphs to attribute causality see Pearl, 2009; Pearl et al., 2016).

It is worth noting that this particular set of rules is compatible with different procedures being inside the box labeled "Look At." This might involve using drill and practice, watching instructional videos (edutainment), reading web pages, natural language tutors, etc. A student might also choose not to review the information and just re-take the test. There are also many options for how to produce a score. Often this will be a percentage correct, but educational measurement experts discuss many alternatives. There are advantages and disadvantages to not constrain the rule specifics, particularly when the system

FIGURE 3 | A single module. The student can go back to look at content and then repeat the test, over and over. This is called a directed cyclical graph.
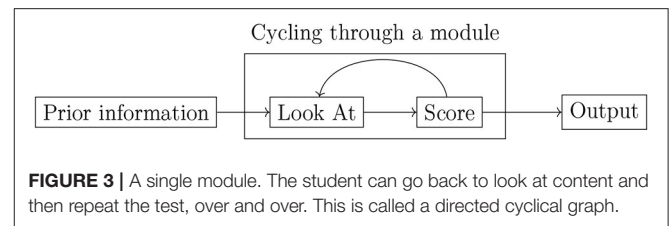
is designed to cover many different academic subjects. If these specifics are not stated in the rules then they can be varied. Evaluations should, however, take into account this variability. The implementation level in **Table 2** lists three possibilities. It is possible to implement the rules without a computer, using for example professional human tutors (peers can sometimes also be used). This might be practical for some tasks, like learning to drive, but this would be prohibitively expensive to educate all students throughout their curricula. This is why most people view computers as the most practical way to implement the rules listed in **Table 2** and thereby to achieve the goals. For this reason PL has become closely associated with EwT.

An Approach to EwT

The purpose of this section is to provide recommendations for two groups of people and some example research questions. The groups are: (a) individual researchers, and (b) people buying EwT products.

Individual Researchers

Figure 3 isolates a small number of relationships so that they can be more easily understood in isolation. This is a common approach to science: examine the phenomenon "in the simplest possible context that is not entirely trivial, and later generalize" (Cox and Donnelly, 2011, p. 5). This allows researchers to identify causal relationships. The difficulty of this approach is that how things operate in complex contexts can be different than how they operate in a simple context. McGuire (1973, p. 452) discusses this as his second Kōan for research: "In this nettle chaos, we discern this pattern, truth." In most naturally occurring situations the variables in which the researchers are interested will likely covary with many other variables. Untangling these relations (and "nettle chaos" seems a good metaphor for having many relationships among variables) is statistically difficult. McGuire describes how this approach requires statistical expertise, but that it should be guided by theory. The following is a list, influenced by McGuire's (1973), for researchers to consider.

1. For researchers focusing just on a small part of the system, like that depicted in **Figure 3**, the relationship between just two variables can be complex. Often the relationships are non-linear and vary depending on so called moderator variables.
2. If your theory predicts that some intervention affects test scores, describe what else it should affect and what it should

not affect. And then measure these. The famous geneticist and statistician, Ronald Fisher, summed this up nicely: "Make your theories elaborate" (see Cochran, 1965, p. 252).

3. If your primary effect of interest is long-term, like graduating high school, it can be useful to include mediator or proxy variables. These are variables that if affected signal that the long term variables should also be affected. For example, with graduating high school, mediator variables would include lower delinquency which has a causal influence on graduation. Proxy variables are common in medicine. If you are interested in whether some drug given to people in their forties reduces the risk of heart attacks later in life, you might measure blood pressure in the weeks after giving the drug to subjects and extrapolate that by reducing blood pressure in the short term this drug also reduces the probability of a heart attack in the long term.

4. When trying to show how a single "nettle plant" works within "nettle chaos" it may be useful to draw all the variables/constructs that you are interested in and use arrows to show how they may be connected. When many variables are all inter-related, trying to understand causal and associative relationships from a complex graph can be difficult. Pearl (2009) is a key reference for identifying causes from graphs (see also Morgan and Winship, 2007; Imbens and Rubin, 2015; Pearl et al., 2016; Steiner et al., 2017). Simulation methods can also be useful to show the predicted outcomes from these diagrams.

5. The focus of much science is on causal relationships, but in some cases associative relationships are also important. Researchers should not confuse these and should use appropriate methods for investigating each of these. It is important to avoid causal words, like "influence," "effect," and "impact," when the study was not designed to estimate causal relations (Wright, 2003).

6. Particularly when studying a complex system it is necessary to have theory guide analyses to avoid data fishing/mining problems. The relationships assumed for the rules and core features should help to constrain the statistical analyses.

Caveat Emptor: What the Buyer Should Know

Modern EwT consumers face the difficult situation when deciding whether to acquire EwT products and if so which ones. One solution would be to have regulations on what can be sold, but this seems unlikely in the current political climate in the US. An alternative is to have consumers demand more verifiable information from vendors before they spend money. If EwT research is done well and consumers expect vendors to provide certain information, then we can move beyond the wild west situation without further regulations. The following list are aspects of the product that ideally a consumer should be told. At present most vendors will not have answers to all of these, but hopefully future research will provide them with answers. This list is based on how judges in the US are told to decide whether to accept expert testimony on scientific and technical matters. These guidelines are adapted from the Federal Rules of Evidence and primarily from three US Supreme Court decisions. These are collectively called the *Daubert Trilogy* (Daubert, 1993; Joiner, 1997; Kumho, 1999).

1. The evidence supporting the product's effectiveness should be generally agreed upon by those in the relevant field (e.g., learning scientists).

2. The studies that provide the evidence should be based on sound scientific principles. *Daubert* discusses Popper's (1959/2002) use of falsification to demarcate science from non-science. According to Popper a good scientific theory should have withstood studies that could have falsified it. If so, it is said to have attained a degree of corroboration. Additional aspects of the scientific value of the supporting evidence should also be considered.

3. The effect sizes (or error rates) of any effects should be known. The vendor should be able to predict effect sizes for your school, should reference the uncertainty of these estimates, and should be able to say how these estimates were calculated. The variables (e.g., school and student variables) that moderate efficacy should be known and the situations where it is not predicted to be effective should be discussed. The intervals for the estimates may be very broad, but the uncertainty of estimates should not be hidden.

4. The evidence that the vendor uses to support their claims should be published in peer-reviewed journals that adhere to scientific principles. It is important for consumers to realize that that being published in a good journal does not imply the finding is accurate. There are many inaccurate findings in good journals (Ioannidis, 2005). However, the peer review process does prevent much junk science from being disseminated.

5. The distance between the research and the conclusions should not be too great. This might relate to how well the rules match with core features and these with goals, whether the conditions used for the supporting evidence are very different from the school setting, or whether the sample in the studies is very different from the intended group of students.

One of the arguments against the *Daubert Trilogy* is that it requires judges to make complex judgments about the value of scientific research when they do not receive much training on this. Differentiating junk science from reputable science can be difficult. Those making IT decisions for school systems are in a similar situation. They may face enthusiastic vendors and need to differentiate circumstance from pomp.

Example Research Questions

A few example research questions are presented to provide a flavor of the type of research that can address whether a product should work. For illustration the following list will focus on the type of EwT depicted **Figure 3**.

Assessment

How does the software estimate whether the student is likely to have reached the desired performance level? Is the assessment fair, valid, and reliable? How do the scores given to students affect how they decide to navigate the system and how do they affect students' beliefs about how much they know?

Bad choices

People learn from making errors (Metcalfe, 2017) and given that many EwT systems require students to make choices, some of these will be bad choices. Can these be identified and types of bad choices classified? Can students be taught to make better choices and are there ways to ensure that students will learn as much as possible from their bad choices?

Analysis

Computer software can provide a large amount of data (i.e., the log files). While exploratory data mining might suggest some associations, the data are messy and exploratory atheoretical mining can be problematic (massive "nettle chaos" with statistical analyses with many researcher degrees of freedom). Analysis based on theories of the students' cognitive processes could direct specific statistical questions.

Agency

How does choosing one's own path affect confidence? Are students happier with the task if they believe that they have chosen it? Are they likely to engage in the task more? Is there less mind-wandering? These could all mediate the effect of agency.

SUMMARY

"Modern technology will dramatically improve education!" attracts headlines, but educators have read headlines like this before. Technology has the *potential* to improve education and it might someday revolutionize education, but to date research evidence has failed to keep pace with optimistic rhetoric. The computer is different than past technologies because students are already learning about computers and many have computers at home (and in their pockets). Even compared with a decade ago, children are more immersed in computing technologies (i.e., many are *Digital Natives*). There are still pitfalls and it is possible (though unlikely) that EwT could fade as some other technologies have. It is important to learn from the history of EwT so that the field does not succumb to the same problems and to understand the current environment so that other potential problems can be addressed. There is much investment both financially and by many schools changing how they teach students, so there are many people wanting this to work. The

goal of this paper was to put forward a research framework to increase the likelihood of success and to maximize positive impacts.

The best way to avoid pitfalls is to accumulate evidence about the effectiveness of products, submit the research for peer-review, and show how continued improvements to products are helping students. It is important to show those contemplating EwT that its adoption is a good investment. This requires more evidence–of the type described in the list for what consumers should ask vendors–to show that using these new technologies is financially responsible.

A research framework was put forward that will help with these goals. The focus should be to show that the different components of the system work. Studies should not just look at whether an innovative program improves end-of-year test scores, but whether the individual parts of this program influence many of the facets that co-vary with and influence test scores. This will help the field to evolve and to show why products work. The focus should not be on specific products, because new versions of them will arrive with new technologies (at the implementation level), but on the rules that the products implement and on whether these rules lead to the goals of the system. It is important for EwT to have evidence to withstand criticism. It is important that researchers and developers continue to strive for Suppes' goal to provide Aristotle-like EwT tutors for all students.

AUTHOR CONTRIBUTIONS

The author confirms being the sole contributor of this work and approved it for publication.

FUNDING

ACKNOWLEDGMENTS

REFERENCES

André, E., Baker, R., Hu, X., Rodrigo, M. T., and du Boulay, B. (eds.). (2017). *Artificial Intelligence in Education*. Gewerbestrasse: Springer.

Arney, L. (2015). *Go Blended! A Handbook for Blending Technology in Schools*. San Francisco, CA: Jossey-Bass.

Atkinson, R. C. (1968). Computerized instruction and the learning process. *Amer. Psychol.* 23, 225–239. doi: 10.1037/h0020791

Benjamin, L. T. Jr. (1988). A history of teaching machines. *Am. Psychol.* 43, 703–712. doi: 10.1037/0003-066X.43.9.703

Borges, J. L. (1941/1998). *Collected Fictions*. New York, NY: Penguin Books.

Carruthers, M. (1990). *The Book of Memory: A Study of Memory in Medieval Culture*. New York, NY: Cambridge University Press.

Clark, A. (2008). *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. New York, NY: Oxford University Press.

Cochran, W. G. (1965). The planning of observational studies of human populations. *J. R. Stat. Soc. Ser. A* 128, 234–266. doi: 10.2307/2344179

Conati, C., Heffernan, N., Mitrovic, A., and Verdejo, M. (eds.). (2015). *Artificial Intelligence in Education*. Gewerbestrasse: Springer.

Cox, D. R., and Donnelly, C. A. (2011). *Principles of Applied Statistics*. Cambridge: Cambridge University Press.

Cuban, L. (1986). *Teachers and Machines: The Classroom use of Technology Since 1920*. New York, NY: Teachers College Press.

Cuban, L. (2001). *Oversold and Underused: Computers in the Classroom*. Cambridge, MA: Harvard University Press.

Daubert (1993). *Daubert v. Merrell Dow Pharmaceuticals, Inc.* 509 US 579, 589.

Deaton, A., and Cartwright, N. (in press). Understanding and misunderstanding randomized controlled trials. *Soc. Sci. Med.* doi: 10.1016/j.socscimed. 2017.12.005

Ferster, B. (2014). *Teaching Machines: Learning From the Intersection of Education and Technology*. Baltimore, MD: Johns Hopkins Press.

Ferster, B. (2016). *Sage on the Screen: Education, Media, and How We Learn*. Baltimore, MD: Johns Hopkins Press.

Foley, B., and Goldstein, H. (2012). *Measuring Success: League Tables in the Public Sector*. London: British Academy.

Ginsberg, A. (1955/2014). *Howl and Other Poems*. Mansfield Centre, CT: Martino Publishing.

Ginsburg, A., and Smith, M. S. (2016). *Do Randomized Controlled Trials Meet the "Gold Standard"? A Study of the Usefulness of RCTS in the What Works Clearinghouse*. Technical report. American Enterprise Institute.

Greenfield, S. (2015). *Mind Change: How Digital Technologies Are Leaving Their Mark on Our Brains*. New York, NY: Random House.

Horn, M. B., and Staker, H. (2015). *Blended: Using Disruptive Innovation to Improve Schools*. San Francisco, CA: Jossey-Bass.

Imbens, G. W., and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York, NY: Cambridge University Press.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med.* 2:e124. doi: 10.1371/journal.pmed.0020124

Joiner (1997). *General Electric Co. v. Joiner*. 522 U.S. 136.

Kumho (1999). *Kumho Tire Co. v. Carmichael*. 526 U.S. 137.

MacKay, C. (1841/2012). *Extraordinary Popular Delusions and the Madness of Crowds*. San Bernardino, CA: Renaissance Classics.

Marr, D. (1982/2010). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Cambridge, MA: MIT Press.

Marr, D., and Poggio, T. (1977). From understanding computation to understanding neural circuitry. *Neurosci. Res. Prog. Bull.* 15, 470–488.

McGuire, W. J. (1973). The yin and yang of progress in social psychology: seven koan. *J. Pers. Soc. Psychol.* 26, 446–456. doi: 10.1037/h0034345

Metcalfe, J. (2002). Is study time allocated selectively to a region of proximal learning? *J. Exp. Psychol. Gen.* 131, 349–363. doi: 10.1037/0096-3445.131. 3.349

Metcalfe, J. (2017). Learning from errors. *Annu. Rev. Psychol.* 68, 465–489. doi: 10.1146/annurev-psych-010416-044022

Morgan, S. L., and Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Cambridge: Cambridge University Press.

Muller, J. Z. (2018). *The Tyranny of Metrics*. Princeton, NJ: Princeton University Press.

Nolen-Hoeksema, S., Fredrickson, B. L., Loftus, G. R., and Lutz, C. (2015). *Atkinson and Hilgard's Introduction to Psychology, 16th Edn*. Wadsworth: Cengage Learning.

Nye, B. D., Graesser, A. C., and Hu, X. (2014). AutoTutor and family: a review of 17 years of natural language tutoring. *Int. J. Artif. Intell. Educ.* 24, 427–469. doi: 10.1007/s40593-014-0029-5

Pane, J. F., Griffin, B., McCaffrey, D. F., and Karam, R. (2014). Effectiveness of cognitive tutor algebra I at scale. *Educ. Eval. Pol. Anal.* 36, 127–144. doi: 10.3102/0162373713507480

Pane, J. F., Steiner, E. D., Baird, M. D., and Hamilton, L. S. (2015). *Continued Progress: Promising Evidence on Personalized Learning*. Technical Report RR-1365-BMGF, RAND Corporation.

Papert, S. (1980). Paper for the president's commission for a national agenda for the 80s. Available online at: http://www.papert.org/articles/president_paper.html

Papert, S. (1993). *The Children's Machine: Rethinking School in the Age of the Computer*. New York, NY: Basic Books.

Papert, S., and Caperton, G. (1999). *Vision for Education: The Caperton-Papert Platform*. St. Louis, MO: Essay written 91st Annual National Governors' Association.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference, 2nd Edn*. New York, NY: Cambridge University Press.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Chichester: Wiley.

Poggio, T. (2012). *The Levels of Understanding Framework, Revised*. Technical Report MIT-CSAIL-TR-2012-014, CBCL-308, MIT, Cambridge, MA.

Popper, K. (1959/2002). *The Logic of Scientific Discovery*. Milton Park: Routledge.

Pressey, S. L. (1933). *Psychology and the New Education*. New York, NY: Harper and Brothers Publishers.

Reingold, J. (2015). Why ed tech is currently 'the wild wild west'. *Fortune*. Available online at: http://fortune.com/2015/11/04/ed-tech-at-fortune-global-forum-2015

Ritter, S., Yudelson, M., Fancsali, S. E., and Berman, S. R. (2016). "Mastery learning works at scale," in *Proceedings of the Third ACM Conference on Learning @ Scale* (Edinburgh: ACM), 71–79.

Roediger, H. L. III. (1980). Memory metaphors in cognitive psychology. *Mem. Cogn.* 8, 231–246.

Roediger, H. L. III., Agarwal, P. K., Kang, S. H. K., and Marsh, E. J. (2010). "Benefits of testing memory: best practices and boundary conditions," in *New Frontiers in Applied Memory*, eds G. M. Davies and D. B. Wright (Brighton: Psychology Press), 13–49.

Rubin, D. C. (1995). *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-out Rhymes*. New York, NY: Oxford University Press.

Sagan, C. (1996). *The Demon-Haunted World: Science as a Candle in the Dark*. New York, NY: Ballantine Books.

Skinner, B. F. (1958). Teaching machines. *Science* 128, 969–977. doi: 10.1126/science.128.3330.969

Small, J. P. (1997). *Wax Tablets of the Mind: Cognitive Studies of Memory and Literacy in Classical Antiquity*. London: Routledge.

Steiner, P. M., Kim, Y., Hall, C. E., and Su, D. (2017). Graphical models for quasi-experimental designs. *Sociol. Methods Res.* 46, 155–188. doi: 10.1177/0049124115582272

Sullivan, G. M. (2011). Getting off the "gold standard": randomized controlled trials and education research. *J. Grad. Med. Educ.* 3, 285–289. doi: 10.4300/JGME-D-11-00147.1

Suppes, P. (1966). The uses of computers in education. *Sci. Am.* 215, 20–220. doi: 10.1038/scientificamerican0966-206

Taylor, R. D., and Gebre, A. (2016). "Teacher–student relationships and personalized learning: implications of person and contextual variables," in *Handbook on Personalized Learning for States, Districts, and Schools*, eds M. Murphy, S. Redding, and J. S. Twyman (Charlotte, NC: Information Age Publishing, Inc.), 205–220.

Thornburg, D. (2014). *From the Campfire to the Holodeck: Creating Engaging and Powerful 21st Century Learning Environments*. San Francisco, CA: Jossey-Bass.

Trotsky, L., (1933). "What is national socialism? [English translation]," in *Trotsky Internet Archive, originally The Modern Thinker*. Available online at: https://www.marxists.org/archive/trotsky/germany/1933/330610.htm

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Boston, MA: Harvard University Press.

Wright, D. B. (2003). Making friends with your data: improving how statistics are conducted and reported. *Brit. J. Educ. Psychol.* 73, 123–136. doi: 10.1348/000709903762869950

Wright, D. B. (2006). Causal and associative hypotheses in psychology: examples from eyewitness testimony research. *Psychol. Public Policy Law* 12, 190–213. doi: 10.1037/1076-8971.12.2.190

Yates, F. A. (1966). *The Art of Memory*. Chicago, IL: University of Chicago Press.