ACTA DE INVESTIGACIÓN PSICOLÓGICA, 2014, 4 (3), 1657 - 1672

Gathering Evidence for Distance Education[1]

Philip N. Chase[2]
University of Massachusetts Medical School

Abstract

A technology for evaluating computer-based distance education curricula for children and people working with children is described. The technology originated from a model of evaluation described by Markle (1967). The components were elaborated through data-based decisions reported in technical reports for a reading acquisition program, two math programs, a curriculum for people with autism, and a professional development program for clinicians working with children and adolescents. The article integrates single-case and group evaluation strategies, and draws attention to the need for better data in evidence-based decisions, and the use of data in continuous improvement efforts. Details concerning the individual learner at the developmental level of evaluation are emphasized, including an illustration of an e-learning rubric assisting this level of evaluation.

Keywords: E-learning, Distance Education, Evidence-based curricula, Evaluation, Single-case Research, Rubrics, Expert Review

Juntando Evidencia para la Educación a Distancia

Resumen

Se describe una tecnología por computadora para evaluar curricula para la educación a distancia para niños y personas que trabajan con niños. La tecnología se originó de un modelo de evaluación descrito por Markle (1967). Los componentes se elaboraron a través de decisiones basadas en datos, publicadas en reportes técnicos sobre un programa de adquisición de la lectura, dos programas de matemáticas, un curriculum para personas con autismo y en el desarrollo de un programa para clínicos que estaban trabajando con niños y adolescentes. El artículo integra estrategias de evaluación de un solo caso y de grupos y hace hincapié en la necesidad de obtener mejores datos para la toma de decisiones basada en evidencia y para el continuo mejoramiento de los esfuerzos. Se enfatizan los detalles relativos al aprendiz individual a un cierto nivel de desarrollo y evaluación, incluyendo una ilustración de una rúbrica de un e-aprendiz asistiendo este nivel de evaluación.

Palabras Clave: E-aprendizaje, Educación a Distancia, Curricula Basada en Evidencia, Investigación de un Solo Caso, Rúbricas, Revisión por Expertos

Original recibido / Original received: 18/06/2014 Aceptado / Accepted: 01/10/2014

[1] The author wishes to acknowledge the following people for their assistance in developing the methods described in this article: Robert Collins, Chata Dickson, Charles Hamad, T.V. Joe Layng, Andrew Lightner, Harold Lobo, Kristin Mayfield, John Rochford, Janet Twyman, and Vennessa Walker. Support for developing these methods came from grant #10009793-1003772R from iLearn, Inc. to West Virginia University; and Interagency Service Agreement # CT EHS 8UMSCANSISA0000001CB between the University of Massachusetts Medical School and the Children's Behavioral Health Initiative of the Executive Office of Health and Human Services, Commonwealth of Massachusetts.

[2] Correspondence: Philip N. Chase, Ph.D., E.K. Shriver Center, University of Massachusetts Medical School, 465 Medford Street, Charlestown, MA 02129

This article describes issues that have arisen while developing a technology for evaluating computer-based distance education curricula for children and people working with children. The article extends a series of editorials I wrote for the *Current Repertoire,* the newsletter for the Cambridge Center for Behavioral Studies between the winter of 2008 and the spring of 2010 (Cambridge Center for Behavioral Studies, 2014). The article also uses data collected and reported in technical reports for a reading acquisition program, two math programs, a curriculum for people with autism, and a professional development program for clinicians working with children and adolescents. My goal is to report on best practices for evaluating e-learning from a behavior analytic perspective.

A New Dawn for Behavior Analysis

A new dawn has risen for behavior analysts. We have a wonderful opportunity to accomplish many things today because so many people are responding positively to our science. Parents, pediatricians, psychologists, and teachers opt for behavioral treatment plans for people with autism and other developmental disabilities. Zoos and pet owners hire behavior analysts to solve significant problems related to human interaction with other species. Managers, front-line supervisors, workers, and unions recognize the importance of behavioral safety. Record numbers of people attending behavioral conferences attest to these positive reactions from the culture at large. These successes have positioned behavior analysis to have an impact on other areas of human concern involving learning, like the development of e-learning or distance education.

I suggest that we should tread carefully. Behavior analysts have squandered their influence on education before. The history of two significant educational innovations by behavior analysts, Programmed Instruction (PI) and the Personalized System of Instruction (PSI) are informative (Bernstein & Chase, 2012). Both PI and PSI were successful for short periods of time in the main culture. Despite the best efforts of researchers and curriculum designers like Donald Cook, Francis Mechner, Susan Markle, James Holland, Beth Sulzer-Azaroff, and others, quality control lapsed, and so did PI. Similarly, despite the work of many who showed repeatedly that PSI was superior to lectures (e.g., Johnson & Ruskin, 1977; Kulik, Kulik, & Cohen, 1979), adopting the structure of PSI without integrating thorough evaluation did not change the modal method of teaching in universities: we still lecture. Even many forms of distance education try to maintain features of the lecture method, e.g., Harvard's HBX Live (Lavelle & Ziomek, 2013).

I submit that our greatest care should come from assuring that we do not give short shrift to quality control: collecting and communicating the evidence behind our successes. One of the primary technologies of behavior analysis is the technology of gathering evidence about behavior change. In what follows, I will address evidence-based practices in the development of curricula. I will describe some of the general strategies with details--the tactics being developed through our work -- that turn practices into technological solutions to curriculum problems. Like most technologies, behavioral technologies are tied to the critical feedback

provided by scientific methods. Without this feedback, the enterprise collapses. Behavior analytic solutions, like those from any field, are only as good as their last evaluation, and evaluation is only as good as the methods used. What are these methods?

Methods of Evaluation

Behavior analysts typically use experimental methods to evaluate their work. An experiment involves the manipulated comparison of a phenomenon under two or more conditions to minimize plausible alternative explanations (internal validity) and test the generality of results across contexts (external validity), while demonstrating reliability of measurement and replication of procedure. An experimental analysis allows investigators to gather strong evidence that can support an educational practice or show that it does not work.

The current standard for experimental analyses used by educators is the random control experiment or trial (RCT). Educators widely accept the RCT because the logical coherence of a random controlled experiment is exceedingly simple to understand. Random sampling from the population suggests that findings will apply to members of the defined population. Dividing a sample into two or more groups and randomly assigning members to the comparison conditions minimizes alternative explanations for the results. One can answer questions concerning the external and internal validity of a particular e-learning curriculum in a few well-designed studies if one uses an RCT strategy effectively.

Random controlled experiments, however, are only one kind of experiment. Discuss the field of developmental disabilities for a nanosecond and one encounters the concept of functional analysis. As an experimental tactic, functional analysis evaluates which consequences are likely to support a problem behavior. In the simplest case, a clinician manipulates one of the purported reinforcers, for example, attention. Across repeated manipulations, if one finds an increased likelihood in a problem behavior when attention is presented contingent on the problem behavior and no increased likelihood in the problem behavior when attention is presented non-contingently on the problem behavior or contingent on another behavior, then the clinician may conclude that attention functions as a reinforcer. The logic of a single-case experiment illustrates its clarity. The repeated manipulation of a variable across time with one individual helps us understand a functional variable for this individual. This logic suggests why single-case experiments became a powerful part of the technology of evaluation for behavior analysts.

Rather than seeing RCTs and single-case experiments as two parts of an experimental strategy to gather evidence, however, behavior analysts and other educational scientists often have butted heads over experimental tactics. “Us vs. Them” arguments have delayed an integrated approach to evaluation. Yes, many bad decisions have been made using inferential statistics poorly to back up the findings of an RCT (e.g., Branch, 1999) and yes, there are practical problems with RCT’s in schools (e.g., the large number of schools, teachers, and students create administrative road blocks). Single-case experiments also have been criticized for

potential subject biases, lack of generality, and lack of standards for evaluating results (Horner & Spaulding, in press). These problems have been addressed and educators have finally agreed on standards for using single-case methods as well as RCT's to evaluate educational practices. (Horner & Spaulding, in press; Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf, & Shadish, 2010). I will not discuss these issues further. Instead I will focus on how a combination of single-case experiments, group experiments, and other forms of evaluation can provide strong evidence of what works for developing e-learning curricula for children and professionals working with children.

The model for curriculum evaluation that my colleagues and I have used as the foundation of our evaluations of curricula, particularly those used in distance education, has been available for many years. Markle (1967) described key components of this evaluation model in a chapter on programmed instruction. Markle described three levels of evaluation that synthesized single-case, small n experiments, and large n group methods. She named these levels developmental, validation, and field-testing. They map well onto current stages of evaluation described by the U.S. Department of Education as development, validation, and scale-up. What follows illustrates these levels of evaluations from our recent work.

Developmental Evaluation. First, intensive individual interactions evaluate the development of an educational practice and document its effectiveness. Markle refers to this level of evaluation as developmental testing. For developmental testing, the curriculum designer/evaluator examines communication problems, learning problems, and motivation problems. Because this level of evaluation has received less attention than single-case and group experimental methods, my colleagues and I designed our own two-pronged tactic. One prong implements a rubric that a trained instructional designer uses to check the accordance of online instruction with the best practices of applied behavior analysis (Bernstein & Chase, 2012), universal design (Universal Design, 2012) and accessibility (WEBAIM, 2012). The other prong of our developmental testing involves frequent interaction with the learner as they progress through the material. We iterate between these tactics while evaluating curricula.

Our use of a rubric for evaluating instructional practices began with selecting tools to assist in writing computer-based instruction (Chase, 1985). Figure 1 illustrates the general characteristics of a rubric that has evolved since 1985. Educators use rubrics to score complex behavior. Rubrics typically involve at least two dimensions, a list of features and a scoring guide. In our case, we developed the rubric to score and track the complex outcome of developing e-learning curricula. The 8 domains of the rubric are listed on the left of Figure 1: Learning and Motivation, Data Collected and Reported, Plain Language and Readability, Use of Updated Technology, Transformability, Multi-modal, Focus and Structure, and Assistive Technology.

Domains	Comments	Absent	Weak	Adequate	Strong	Excellent
1. Learning and motivation						
2. Data collected and reported						
3. Plain language and readability						
4. Updated use of technology						
5. Transformability						
6. Multi-modal						
7. Focus and structure						
8. Assistive technology						

Figure 1. Instructional design rubric with general domains as rows and comments and ratings as columns.

Each domain on the rubric can be commented upon and rated on the five-point scale listed on the top of the tool. We also expand each row or domain on the rubric to a set of features and rate and comment on each of these features. Figure 2 shows a representative subset of the 17 features of the Learning and Motivation domain**.** The features of Learning and Motivation come from a very strong tradition of experimental evidence. As we audit the instruction we ask: Does the instruction provide sufficient examples to test for discrimination between classes and generalization within classes of stimuli? Does the instruction include immediate, frequent, and differential consequences?

1. Learning and motivation	Comments	Absent	Weak	Adequate	Strong	Excellent
8. Mastery requirements						
12. Discrimination among classes assessed						
13. Generalization among classes assessed						
14. Immediate consequences						
15. Frequent consequences						
16. Differential consequences						

Figure 2. Instructional design rubric: Representative features of the Learning and Motivation domain with illustrative features listed in the rows, and comments and ratings in the columns.

We used the rubric as part of an evaluation of Headsprout.com, an English language reading acquisition program. My colleagues and I conducted an expert or peer review of Headsprout as well as an experimental evaluation of it in two kindergarten classes (Walker & Chase, 2006). We also evaluate two math curricula from iLearn.com: iPass and Thinkfast (Chase, Dickson, Alligood, Lobo, Walker, & Cook, 2007; Chase, Dickson, Alligood, Lobo, & Walker, 2008). Again our evaluation included a review using a version of the rubric and an experimental analysis in our lab with children from the community. A team of experts also evaluated the Autism Curriculum Encyclopedia® (ACE) curriculum from the New England Center for Children using a version of the rubric (Chase, Alai-Rosales, Smith, & Twyman, 2012). ACE is a web-based toolkit providing special educators with an evidence-based program to effectively assess, teach, and evaluate individuals with autism. Most recently we have used the rubric to review the Child and Adolescent Needs and Strength (CANS) Training program for the state of Massachusetts (Bondardi, Chase, Hall, Lauer, & Nubrett-Dutra, 2013). I will use our evaluation of CANS to illustrate the tool.

CANS is a communication and care coordination instrument. It supports decision-making, facilitates quality improvement initiatives, and helps monitor the outcomes of behavioral health services for children and youth. Any clinician who provides behavioral health care to a client under the age of 21 and receives funding from Mass Health in Massachusetts is required to use CANS. Mass Health is the public health insurance program for low- to medium-income residents of Massachusetts.

Clinicians must be certified to use CANS with clients. Our evaluation focused on the e-learning certification training designed by a team from the University of Massachusetts Medical School. Our review required frequent iteration between the training and testing materials and the features of the rubric. We surveyed as many components of the training as possible from beginning to end. Then we developed questions to examine various features of the training. We returned to the beginning of the training and read, watch, and listened to each component, attempting to answer questions generated from the survey as well as creating further questions.

During the audit, the features prompted by the tool were noted qualitatively, rated, and the notes and ratings became the substance of the review. The rubric was examined frequently to assure a thorough review of all its features. For example, one of the accessibility features included checking for delivery in multiple modalities. As we examined the training we questioned whether critical components included text, audio, and rich media. We tested these features. And then we completed the comments and rated them before moving on to other features that were being checked.

The review for the domain of multi-modal is shown in Figure 3 for the CANS certification training.

6. Multi-modal	Comments	Absent	Weak	Adequate	Strong	Excellent
1. Content in multiple mediums	Videos not included for some critical learning			x		
2. Video and audio alternatives	Alternatives throughout, though not inspirational				x	
3. Text alternatives	A little confusing in placement				x	
4. Closed captioning						x
5. Illustration, diagrams, icons, and animations used to convey complex information	Inconsistent in placement			x		
6. Pair icons, graphics, etc. with text	Some alerts				x	

Figure 3. Instructional design rubric with the Multi-modal domain completed for CANS certification training.

A sample of the strengths and recommendations indicating how the rubric is translated into a review (Bondardi, Chase, et al., 2013) is provided below:

> A simple, dignified, no-nonsense, well-designed interface allows a straightforward navigation through the materials. Simple language use with jargon and abbreviations kept to a minimum help the leaner understand the material. Intermittent tasks for learners are provided to check their understanding with clear, immediate, and frequent feedback on their responses. Training ends with full case practice examples (vignettes) that helps integrate learning. In addition, multi-modal training is used throughout with closed captions and transcripts for voice and videos that helped focus training on some of the more difficult domains and items within domains.
>
> As currently designed, however, the training and testing do not provide sufficient interactions with a range of examples to teach discriminations between some of the most difficult items and generalization within these items. Further, if the learner responds incorrectly to the questions provided, there is no chance for the learner to recheck learning with a new question on the same item. (Bondardi, Chase, et al., p. 5).

In sum, the rubric helps set standards for peer/expert reviews of online /distance education. It prompts us to examine various features of the curriculum and how it is presented online. It helps to make reviews and critiques efficient. Most importantly, it synthesizes what we know from behavioral education with what we know from accessibility into one set of standards that we can apply to online instruction and training.

Peer or expert review, however, is not sufficient even when standardized as we have done. Like the problems with computer software that arise when checked and tested only by software engineers, end-users (typical learners) should test educational programs. The end-user evaluation we conducted for the iPASS math curriculum illustrates how we interacted with the students as they progressed through the curriculum. iPASS is a web-enabled mathematics curriculum for middle-school students used in several states in the US (iLearn.com, 2014). Teachers and students use iPASS as either a primary or supplementary source of mathematics instruction. The software automates many aspects of the instructional process, including placement, assessment, instruction, remediation, and tracking of student performance.

In addition to an expert team reviewing and critiquing iPASS using the rubric, students of iPASS completed a report-on-problems that asked them to detail anything they found to be problematic or frustrating with the iPASS program. The form prompted the student to describe problems and identify where in the curriculum the problem occurred. It also asked them for their level of agreement with a general question about their understanding of the exercises they had completed. The form was place on the desk next to their computer key-board and they were asked to complete the form as they worked through iPASS. At the end of each session, experimenters questioned the students further and then summarized

these reports into a spreadsheet. Part of the students' payment contingency was to suggest changes. Along with problems the evaluators found during their expert review and from student performance data, reports from students were included in the developmental review of the iPASS curriculum (Chase et al., 2007).

Validation Evaluation. One problem with the developmental level evaluation described here concerns basing evidence simply on a combination of expert opinion and simple data from a few students. Such comparisons hardly minimize alternative explanations for the results and they certainly do not provide confidence in external validity. They do not answer questions such as: Could the results be an artifact of something else? Plausible alternatives can be eliminated using Markle's second level of evaluation: the validation stage. Validation testing investigates the extent to which the curriculum meets its own goals under controlled circumstances. Validation testing evaluates internal validity. Some questions concerning external validity can be answered as well, but these questions will be discussed in more detail in the third or field-testing stage. Our validation stage evaluation starts with an experimental analysis.

For the purposes of internal validity we need to control for alternative explanations. One experimental strategy we have used is the single-case design called a Multiple Baseline Achievement Test design (MBLAT) (Miller & Weaver, 1972). It is an example of a multiple-baseline across behaviors design. As such it involves frequent assessment of the multiple behaviors of at least one student changing from baseline to treatment to minimize alternative explanations. Repeatedly examining baseline to treatment changes with baselines of different lengths of time, treatments of different lengths of time, and replication across behaviors, assesses the contribution of the treatment. The treatment in our use of the MBLAT is exposure to an e-learning curriculum. The behavior of individuals in the experiment is the target of evaluation. The changes examined include those of level, trend, and variability of performance across phases.

Figure 4 provides an example with hypothetical data using iPASS as the example of a treatment. Parallel or identical tests are given repeatedly over time (the successive tests are displayed on the x-axis in Figure 5) and each test assesses the same material from a curriculum or a component of a curriculum. In a hypothetical case, the component of the curriculum consists of three math units: one on addition, one on multiplication, and one on word problems involving addition and multiplication. We also divide the test into sections of items that are aligned to the units of the course. The y-axis records the dependent variable (e.g., % correct on each section of the test). The baselines refer to performances on items aligned with each unit **before** the lessons for a unit are provided. The treatment refers to performances on items aligned with each unit **after** treatment (e.g., iPASS math instruction). If training is effective and the test items for each unit are independent of each other, then one should see changes in test performance related to each unit only when the unit material has been taught (post treatment). As illustrated, the changes evidenced were clear changes in level as percent correct performance always increased after instruction and never before. Changes in trend and variability across phases were not evident.

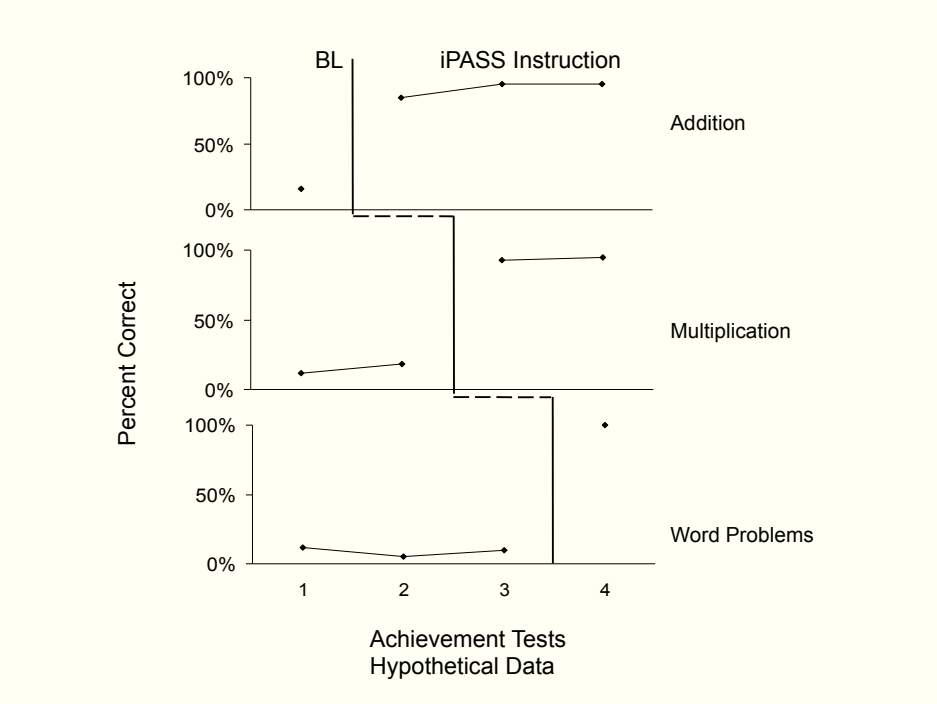

Figure 4. Hypothetical MBLAT across four achievement tests with iPASS Instruction as the treatment, percent correct as the dependent measure, and solving three types of math problems as the behavior.

What are the advantages of the MBLAT design for establishing the validity of a curriculum? First, it is a practical design. Behavioral educators may use the design with many different kinds of curriculum, at many different times in the school year, and with many students or group of students. The design does not depend on random assignment to draw conclusions about the internal validity of the evidence and minimizes the practical problems of using random assignment in ongoing school environments. Students may be randomly selected to participate in the experiment, which strengthens the conclusions one can draw from the MBLAT.

The following list of threats to internal validity adapted from Kazdin (2003) can be examined with the MBLAT: subjects, history, maturation, attrition, selection biases, settings, measurement, instruments, and the adequacy of independent variable (IV) descriptions and definitions (e.g., special treatment in experimental vs. control conditions or diffusion of treatment across conditions). A threat is any known variable that co-varies with the treatment and thus could be a plausible alternative explanation for any changes seen in behavior. To evaluate threats we ask: do subjects, settings, etc., co-vary with changes in conditions.

Many threats to internal validity related to the participants or subjects can be examined with the MBLAT design because the subjects receive both baseline and treatment conditions. If the participant's history concurrent with and outside of the experiment has an impact on their behavior, then it can be evaluated by the staggered introduction of the treatment. Although often described differently,

biological maturation can be described as a plausible change in history. Training research often involves repeated measures over long time periods, and therefore, the participants may experience maturational changes during their time in the experiment that affect learning. These maturational changes can be assessed by the staggered introduction of the treatment. For example, IPASS is a year-long curriculum for children between the ages of 10 and 15, and biological maturation could conceivably affect their behavior. The staggered baseline and treatment phases of the MBLAT allow the researcher to examine whether maturation, the repeated measures themselves, or other variables that occur over the history of the experiment might affect performance. If such historic variables have an impact, the effect would be seen at times other than the introduction of the treatment.

Historic differences among participants prior to the experiment also do not co-vary with the treatment and therefore these threats to internal validity are handled directly by the MBLAT design. If the participants' history with math prior to the experiment allowed them to perform well on the tests, the baseline conditions reveal this. We then assess further changes after treatment. Although differences in baseline performance often occur among participants attributed to historic variables, it is the change in level, trend, and variability of performance from baseline to treatment, replicated across sets of behavior that the MBLAT helps us examine.

Other subject threats, like attrition, frequently pose problems for educational research. Students leave school, transfer classes, and move from school-to-school, so when evaluators use a group design they have to assure that attrition for students who receive a treatment does not differ from those who receive a comparison condition. The MBLAT manages attrition threats again by having each participant as his or her own control. If they leave the experiment it might be costly to the evaluation, but one cannot attribute treatment effects to differential attrition. Attrition affects on external validity are not handled by the MBLAT design as will be described below. Additionally, whether the treatment contributes to attrition cannot be assessed within a participant using the MBLAT, but can be assessed across participants. We can examine attrition as a dependent variable as we add participants: what proportion of the participants leave the experiment during baseline compared to those who leave during treatment conditions?

Other threats such as setting, measurement, and independent variable definitions are handled by assuring that each subject receives the same tests, in the same settings, across all conditions. For example, iPASS uses computers to present the curriculum, therefore, we used computers during baseline instruction. We also used the same tests during baseline and treatment, the same computers were used for testing, and the same people administered testing across conditions. Diffusion between conditions exemplifies problems related to the definition of the independent variable if the baseline and treatment conditions are too similar to each other or influence each other. Special treatment creates another problem with the independent variable if the teachers pay more attention to the kids during treatment than they do during baseline. Controls for threats like test, setting, and IV are managed as in any carefully designed research by the use of highly specified

protocols, and data collected on whether the protocols are carried out as planned and reported as treatment integrity data.

Selection bias is a special case of historic variability. Selection bias refers to the possibility that participants selected for the study have characteristics that make it more likely that they will be affected by the treatment. Selection bias is an identifiable characteristic(s) of the students selected for the study that contributes to the effect of the curriculum. Have we selected students whose special histories allow them to do well in the curriculum? For example, social economic status (SES) factors may influence how the motivation components of a curriculum work. If the students in a study all come from financially privileged families, they may be affected by motivational variables differently than children from less financially advantaged families. Evaluators manage selection bias in RCTs, like all historical variables, by random assignment to conditions. Evaluators manage selection bias in single-case studies because the bias does not co-vary with the treatment. We select the students for the experiment and then we test them under both baseline and treatment conditions. Like other subject threats, selection bias cannot be attributed to the subjects used in the experiment, they are who they are, and if the treatment successfully changes their behavior we should see changes in level, trend, and/or variability. Again, the level of evaluation is individual behavior.

Selection bias, however, does require a little more discussion, a discussion that highlights one aspect of history that MBLAT designs do not eliminate: the interaction of variables with a treatment that could affect efficacy. Selection bias can be a threat if characteristics of the participants interact with the treatment to produce the effects found. But this is a problem of external validity true for almost all the threats I have discussed so far. For example, have we assured that the tests used are not biased toward the curriculum-for example, the problem of teaching to the test? Likewise, have we assured that characteristics of the tested students did not influence the results? As described earlier, randomly selecting students from a population would help to minimize the interaction of selection with the treatment even in a single-case design. In general, however, threats concerning interactions will be considered next under problems of external validity. The MBLAT design allows for a good evaluation of the internal validity for the children who were in the experiment as long as the children, settings, instruments, and tests do not change at the same time as the treatment. The MBLAT does not necessarily evaluate whether these changes will occur across children, settings, instruments, and tests.

Field Evaluation. External validity questions concern whether our results hold up across students, schools, teachers, tests, and other characteristics of the study. Does the study draw conclusions about other students? If so, has the experimenter arranged to test the curriculum with representative participants? Does the study draw conclusions about other environments? If so, has the evaluator arranged to test the curriculum with a representative range of environments? Does the study draw conclusions about the generality to other teachers or staff? If so, has the experimenter arranged to test the curriculum with representative teachers? Does the study draw conclusions about other tests/measures than used? If so, has the evaluator arranged to test whether the curriculum is successful with different measures?

Markle (1967) described how questions of external validity are answered with a field test. She stated that evaluations should be conducted to assess the effectiveness of the curriculum in a variety of settings and with a variety of students. More recently descriptions of such evaluations state that evaluators test the curriculum at scale, the process of "scaling up". A series of MBLAT experiments can be designed to test various questions of generality. One of the most important questions, because of the single-case nature of the MBLAT, is whether the participants are representative of the population that might use the curriculum. A series of such studies may not be practical, however. For example, one practical problem with single-case designs is the difficulty of examining interactions, so even the combination of highly controlled studies to establish internal validity and data collection from representative samples of students, settings, and teachers to establish external validity, may not be sufficient. Once we have established internal validity with a few well-designed MBLAT evaluations, it might be more efficient to use group designs to test for external validity. Discussion of the appropriate designs to use is beyond the scope of this article, but various sources including the IES What Works Clearing House (http://ies.ed.gov/ncee/wwc/) provide useful guidelines.

One important question of external validity has arisen from our work on making distance education accessible to the widest group of students. There are many websites, curricula, and other forms of e-learning that are not accessible to a large population of people. A highly influential medical information website that we examined has flashing ads, pop-ups, cycling banners, multiple columns, and packed information all of which make it difficult for people to access the information. Using the WEBAIM Wave Program (WEBAIM, 2012), which is design to detect accessibility problems primarily related to visual difficulties, we found 17 accessibility errors on the home page of this website. While this site tries to provide a good service, important audiences cannot use it. Who am I talking about? The following list adapted from the Web Accessibility Initiative (Eichner & Dullabh, 2007) is a good start: A mother with color blindness who seeks information for her child with autism, a reporter with repetitive stress injury, an accountant with blindness, a classroom student with attention deficit hyperactivity disorder and dyslexia, a retiree with low vision, hand tremor, and mild short-term memory loss, or a supermarket assistant with Down syndrome.

According to the U.S. Census figures for 2000, 20% of Americans have a disability that impairs access to websites and Internet content. According to a 2011 report on disability from the World Health Organization, 56 million people in the U.S. were identified as having such disability. Multiply these numbers x-fold for a worldwide population of people with intellectual, cognitive, visual, and age-related disabilities who cannot access information and instruction from the internet. These people need access to online instruction and information. How can we design and evaluate online instruction that works for them? The features of evaluation that I have described throughout this article help. We directly address many issues of accessibility with the rubric used during developmental testing. We further address accessibility through field-testing. We collect data across students, across schools, and across tests for a particular curriculum, assuring that we also have evaluated

e-learning with students across critical demographic groups. Once we do so we should have sufficient evidence to establish the external validity of the curriculum's effectiveness. If we combine these data with the data from our MBLAT and the data from our developmental testing, and all of these levels of evaluation demonstrate the effectiveness of a curriculum, do we need additional evidence?

I always return to what I learned from methodologists. Have we addressed alternate explanations, issues of generality, and the practical concerns of gathering evidence? I think we can check these off if we have data that show a curriculum to be effective across students, settings, and tests as well as having internal validity from prior experimentation. If the data show that the curriculum is not effective with some particular set of students, or in some settings, or on some type of test or outcome, then the data suggest further experimentation with the variables correlated with lack of success.

But even if we do achieve internal and external validity in the manner suggested, some might still ask whether the results could be achieved faster or more reliably with another curriculum. This is a consumer driven question. A series of MBLAT studies comparing curricula across phases and with a counterbalance of the order of receiving the two curricula across students can be used to address such questions. The counterbalancing of order has some practical problems related to aligning curricula with tests, but it can be done in many situations. Of course, another solution would be to conduct a random controlled experiment that focuses on comparing the curricula of interest.

Conclusion

The idea of a major industrial concern turning out 10 percent superior products, 20 percent good products, and 50 percent average products, with the remainder classified as disposable is so ludicrous. Markle, (1967), p.104

This quote struck me when I first saw it and it still rings true today for most of what passes for educational technology: the typical measures of success of educational enterprises are absurd for those interested in replicable procedures. Why do we continue to accept them? As indicated at the beginning, Programmed instruction and PSI both made progress. I am humbled when I read a classic like Glaser (1965) or a biography, like Mechner (in press) about how much was done and how many people were educated through Programmed Instruction. So why have we not made more progress since then?

It comes down to what behavior analysts know best: whether contingencies of reinforcement support behavior required of progress. Educators have not had to demonstrate a high level of effectiveness in order to obtain reinforcers. Unlike building bridges across rivers, unlike producing computers that process data more efficiently, unlike reducing pain, teaching has had conflicting goals and uncertain outcomes. Agreed upon demands have not been placed on education to develop thorough evidence-based methods like those from other technologies.

So where do demands on evidence-based education come from? We know how demands for the behavioral services I described at the beginning of this article drove our successes. We know that the changes in the services for children with

autism and other developmental disabilities came from parents' demands. We know that changes in safety practices in industry came from the costs of injuries. We know that changes in zoo animal and pet training came from the consumers. Recently the demands on curricula in the U.S. have begun to change with the attempts to enforce standards through the No Child Left Behind Act (NCLB). The NCLB has many weaknesses. Particularly weak are those features related to the oversimplification of measurement, like standardized test performance. The use of these simple measures in a compliance model has created demands on the wrong behavior by teachers and administrators. I fear these weaknesses once again will derail attempts at data driven change. One of the more promising outcomes of NCLB, however, was the creation of the Institute of Education Sciences (IES) and IES mandates experimental validation of educational practices. For the first eight years this demand translated into using RCT as the gold standard for evidence, but IES has begun to accept single-case experimental designs (Kratochwill, Hitchcock, et al., 2010). A demand on educators in the U.S. to use RCT and single-case experiments to bolster other forms of evidence for what works in education seems to support the kinds of activities I have described here. The question remains whether behavior analysts can help meet and contribute to this demand. I think the technology of evaluation that I have discussed illustrates one behavioral technology that might help.

References

Bernstein, D. J., & Chase, P. N. (2012). Contributions of behavior analysis to higher education. In G. Madden (Ed.), *APA handbook of behavior analysis: Volume 2 Translating principles into practice* (pp. 523-543). Washington, D.C.: APA Press.

Bonardi, A., Chase, P. N., Hall, G., Lauer, E., & Nubrett-Dutra, C. (2013). *CANS training evaluation.* Technical report prepared for Interagency Service Agreement # CT EHS 8UMSCANSISA0000001CB between the University of Massachusetts Medical School and the Children's Behavioral Health Initiative of the Executive Office of Health and Human Services, Commonwealth of Massachusetts.

Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst, 22,* 87-92.

Cambridge Center for Behavioral Studies (October 23, 2014). *The Current Repertoire.* Retrieved from http://behavior.org/repertoire.php.

Chase, P. N. (1985). Designing courseware: Prompts from behavioral instruction. *The Behavior Analyst, 8,* 65-76.

Chase, P. N., Alai-Rosales, S., Smith, T., & Twyman, J. S. (2012). *ACE expert review*. Technical report prepared for The New England Center for Children of its Autism Curriculum Encyclopedia® (ACE).

Chase, P. N., Dickson, C., Alligood, C., Lobo, H., Walker, V., & Cook, L. (2007). *Expert review of iLearn math*. Technical Report 1 from grant #10009793-1003772R from iLearn, Inc. to West Virginia University.

Chase, P. N., Dickson, C., Alligood, C., Lobo, H., & Walker (2008). *Laboratory evaluation of iLearn math*. Technical Report 2 from grant #10009793-1003772R from iLearn, Inc. to West Virginia University.

Eichner, J., & Dullabh, P. (2007). *Accessible health information technology (health IT) for populations with limited literacy: a guide for developers and purchasers of IT.* (Prepared by the National Opinion Research Center for the National Resource Center for Health IT). AHRQ Publication No. 08-0010-EF. Rockville, MD: Agency for Healthcare Research and Quality.

Glaser, R. (1965). *Teaching Machines and programed learning II: data and directions.* Washington: D.C. National Education Association.

Horner, R. & Spaulding, S. (in press). Single-case research designs. *Encyclopedia*. Springer.

iLearn.com (October, 2014). *iLearn.* Retrieved from http://ilearn.com.

Johnson, K. R. & Ruskin, R. S. (1977). *Behavioral Instruction: An Evaluative Review*. Washington, D.C.: American Psychological Association.

Kazdin, A. E. (2003). *Research design in clinical psychology* (4th edition). Needham Heights: Allyn & Bacon.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kulik, J. A., Kulik, C-L. C, & Cohen, P. A. (1979). A meta-analysis of outcome studies of Keller's Personalized System of Instruction. *American Psychologist, 34,* 307-318.

Lavelle, L. & Ziomek, E. (2013). *Harvard business school launching online learning initiative*. Retrieved from http://www.businessweek.com/articles/2013-10-09/harvard-business-school-launching-online-learning-initiative.

Markle, S. M. (1967). Empirical testing of programs. In P. C. Lange (Ed.), *Programmed instruction* (pp. 104–138). Chicago: University of Chicago Press.

Mechner, F. (in press). *Behavioral technology and education reform*. Beverly, MA: Cambridge Center for Behavioral Studies.

Miller, L. K. & Weaver, F. H. (1972). A multiple baseline achievement test. In G. Semb (Ed.) *Behavior analysis and education-1972* (pp. 393-399). Lawrence, KS: Support and Development Center for Follow Through, Department of Human Development, University of Kansas.

Universal Design (August 28, 2012). *The Principles of universal design.* Retrieved from http://en.wikipedia.org/wiki/Universal_design.

Walker, V. L. & Chase, P. N. (2006). *Assessment of Headsprout Basics Reading*. Technical Report for Headsprout.com.

WEBAIM, *Evaluating cognitive web accessibility with WAVE* (August 28, 2012). Retrieved from http://webaim.org/articles/evaluatingcognitive/.