

Scientific Formative Evaluation: The Role of Individual Learners in Generating and Predicting Successful Educational Outcomes

T.V. Joe Layng, Greg Stikeleather, Janet S. Twyman

Headsprout

The effort to bring scientific verification to the development and testing of educational products and practices has begun in earnest (see for example the No Child Left Behind Act of 2001). The first products targeted for improvement are those whose primary goal is the teaching of reading (U.S. Department of Education, n.d.). Other products and practices are sure to follow, particularly if the initial effort is successful in having a major impact on children's reading performance. But what does it mean to take a scientific approach to instructional productivity? This chapter hopes to contribute to that discussion by examining the role scientific assessment can play in enhancing educational productivity through the application of a thorough scientific evaluation *during* the development of instructional programs as well as in their post-development validation. Since reading is the current focus, we shall begin there.

For a beginning reading program to be successful, the National Reading Panel (2000), among others, has identified critical program constituents that scientific investigation has determined as essential to that success. These constituents include: phonemic awareness, phonics, vocabulary,

fluency, and comprehension. Further, it may be argued that the instructional program itself, rather than just its constituents, be research-based. But, what does it mean to be research-based?

RESEARCH BASED: SOME DEFINITIONS

Current uses of the term "research-based" as applied to early reading programs range from claims that (1) programs contain elements that research suggests are effective (see for example Simmons & Kame'enui, 2003), or (2) pretest vs. posttest or simple comparison studies have provided some evidence of effectiveness for an instructional program, or (3) the program has undergone some form of scientifically controlled study, often involving randomized control groups (Coalition for Evidence-Based Policy, 2003). All of these uses, however, fail to distinguish between the *scientific development* of a program, and the *scientific evaluation of outcomes* after a program has been developed.

This latter use of research-based might more properly be considered "research-filtered." That is, a program regardless of how it was designed, is measured against an

alternative form of instruction, or at times, no instruction at all. This use of the term research-based would find in its referent an emphasis on *summative* evaluation (Bloom, Hastings, & Madhaus, 1971). In the research-filtered approach, there is no requirement that the program itself must be scientifically designed or based on research.

Yet another use of the term research-based might be more properly considered as “research-guided.” By that, we are referring to a program of instruction that has been scientifically designed and tested during its development, or at least its design is guided by previous research results. This use of the term research-based would find in its referent an emphasis on *formative* evaluation (Bloom et al, 1971; Scriven, 1974). In the research-guided approach, formative evaluation is intertwined into the instructional design protocols, and at its most thorough influences program development through iterations of testing, revising, and testing again.

LEVELS OF VERIFICATION

Both formative and summative evaluation may comprise varying degrees of verification and commitment to a scientific approach. In the more thorough forms of formative evaluation (also referred to as developmental testing, see Markle, 1967), data are continuously collected and analyzed as the program is developed in order to provide an ongoing *experimentally-controlled* research base for ensuring program effectiveness with individual learners (cf. Sidman & Stoddard, 1966; Stikeleather & Sidman, 1990). In the more thorough forms of summative evaluation, data from randomized, or matched, experimental and control groups are collected and analyzed to provide a *statistically-controlled* research base for determining program effectiveness with groups of learners (Habicht, Victora, & Vaughan, 1999).

In the least thorough forms of formative and summative evaluation, philosophy, point of view, and anecdotal evidence comprise the approach. Little attention is paid to direct measurement of instructional effect, or to the determination of functional relations among variables. Of course, both forms of evaluation also have middle grounds where there is an attempt to use some form of empirical evidence, to influence program development in the case of formative evaluation, and to make judgments about outcomes in the case of summative evaluation.

Table 1 provides a 3 X 2 matrix which describes critical features of formative and summative evaluation across three levels of analysis with each level representing a scale of verification from least to most thorough: experiential evaluation, evidence-based evaluation, and scientific evaluation. Table 2 provides a 3 X 3 matrix depicting the relations between each level of verification as they intersect with one another. The rows denote formative evaluation; the columns denote summative evaluation. The outcome of each interaction across rows and column is described in the cell for each intersection. The level of verification for formative evaluation is indicated numerically, with 1 representing the least thorough, and 3 indicating the most. The level of verification for summative evaluation is indicated alphabetically, with A representing the least thorough, and C indicating the most.

Social validity, which is important for both formative and summative evaluation, is not indicated in the 3 X 3 matrix presented in Table 2. For entries falling in column C, the critical question is, do the measurement instruments chosen for the evaluation reflect what the community considers important? For entries falling in row 3, the critical question is, do the mastery criteria, which each individual performance must meet, reflect what the community considers important? The

discussion of these questions, although of considerable importance, falls beyond the scope of this chapter.

**FORMATIVE EVALUATION:
IMPLICATIONS FOR RESEARCH-
BASED INSTRUCTION**

As noted in the 3 X 3 matrix, programs that evolve from a thorough formative evaluation process can predict *individual* learner outcomes across all summative evaluation levels of verification, just as programs tested under the most thorough form of summative evaluation can predict *group* outcomes across all formative evaluation levels. Both should be considered to have equal predictive power: formative evaluation for individual learners, and summative evaluation for groups of learners. Both formative evaluation (*research-guided* instruction) and summative evaluation (*research-filtered* instruction) are important, and may be combined to provide useful information on individual performance and group averages. Cell 3C of the 3 X 3 matrix represents the ideal intersection between scientific formative evaluation and scientific summative evaluation.

At its most thorough, formative evaluation requires a careful control analysis design to ensure that each constituent part of the program is working alone, or together with other constituent parts, to produce a reliable and predictable outcome (Goldiamond & Thompson, 1967, reprinted 2004). Accordingly, such formative evaluation lends itself most readily to single-subject research designs (Bernard, 1865, [translated 1927]; Neuman & McCormick, 1995; 2002; Sidman, 1960) in which participants respond over long periods of time while variables are experimentally changed and controlled. In these designs, variance is controlled through direct procedural or experimental intervention during the course of

the experiment.

Whereas group designs, typically the basis for summative evaluation, are readily known and accepted as providing scientific evidence for program effectiveness (see Paik, this volume), single subject designs, which are typically the basis for formative evaluation, are not so well known. While both group and single subject designs are descended from highly successful scientific traditions, and both may provide equally thorough and informative results, single subject designs are relatively less understood. Both do, however, differ in the questions asked: one asks about the behavior of groups, the other asks about the behavior of individuals.

**SINGLE SUBJECT CONTROL-
ANALYSIS RESEARCH &
EVALUATION**

Single subject designs are most valuable when the questions addressed concern how program components working alone or together affect an individual's performance. These designs provide predictions on how *individuals* using the program will perform compared to a *standard*; group designs provide predictions on how one *group* will perform as compared to another *group*. Single subject and group designs may also differ in the way variance is typically controlled. In the case of group experimental designs, statistical control and analysis is the primary method of controlling variance, often employing randomized or matched control groups. On the other hand, in single subject experimental designs, procedural change is the preferred method of analyzing and controlling variance.

Although sharing the goal of predicting program outcomes with summative evaluation, the procedural control-analysis designs, which typify formative evaluation, differ from that of summative evaluation and statistical control designs in another important

aspect. Within a procedural control–analysis framework the “...concern is not merely with specifying the conditions under which behavior is appropriate, but with being able to develop such appropriateness, to develop corrective procedures where the existent functional repertoire is inappropriate, and to maintain appropriate functional repertoires once they are developed.” (Goldiamond & Thompson, 1967; reprinted 2004, emphasis added).

Stated differently, in single subject research designs important to formative evaluation, the essential question is whether experimental control is maintained over the learner’s behavior as response criteria are systematically changed (Layng, 1995; also, see for example, Layng, Twyman, & Stikeleather, 2003; Layng, Twyman, & Stikeleather, 2004a; Twyman, Layng, Stikeleather, & Hobbins, 2004). Further, once such control can be demonstrated for an individual, the question is raised as to whether that control can be replicated for other individuals across a range of settings. In such systematic replication (Sidman, 1960), the occurrence of increased variance in responding, both within a learner’s individual performance and between the performance of different learners, is an occasion for the examination of the program elements and sequence in which the variance occurred, and the modification of, or the design of new, procedures so as to reduce or control the variance found in meeting the mastery criteria.

Systematic replication with new individuals provides increased confidence that the same procedures will provide similar outcomes for other individuals. Each new learner can be considered an experimental replication; that is, one can predict that future learners will show similar results. There are many single subject designs that can be used to address different experimental questions.

Regrettably, a detailed discussion of single–subject design is beyond the scope of this chapter. Interested readers are referred to many fine texts on this topic (e.g., Bernard, 1865; Johnston & Pennypacker, 1993; Neuman & McCormick, 1995, 2002; Sidman, 1960).

WHY FORMATIVE EVALUATION AND ITS EMPHASIS ON THE INDIVIDUAL IS SO IMPORTANT

Scientists and engineers whose responsibility it is to design and build working complex systems, such as airplanes, rely on thorough formative evaluation to produce a vehicle that will fly the very first time it is tested. In the case of an airplane, careful wind tunnel and other experiments test how the bolts applied stand up to stress, how the materials used perform, test the lift provided by the wings, and how the overall aerodynamics are implemented. Each revision based on the testing is itself retested until the component meets a quality standard. Only after thorough testing of the components, both separately and together, is the final question asked, “Does it fly?”

Each flight is considered a replication; the more conditions encountered the more systematic the replication. Design modifications that come from test flights serve to improve stability and reliability even more. Aircraft are not constructed and then compared to other aircraft to determine, on average, if one group stays aloft longer than differently built aircraft comprising a control group. Comparative tests between only one or two competing aircraft typically provide enough data for the intended customer to make a buying decision.

Similarly, thorough formative evaluation may have the same effect on teaching reading and other instructional program development (see for example Markle & Tiemann, 1967; Twyman, et al, 2004). By ensuring that each

component meets a specified quality standard, which in the case of instruction would be a high mastery standard achieved by the learners tested, we should be able to design and build instructional programs that have the same high likelihood of success as does building modern aircraft. Rigorous “single–subject” iterative cycles (test–revise–test) provide great confidence that all aircraft built in accord with the design and development process will fly—without the need for tests comparing groups of aircraft. A similar approach to educational program development may provide comparable confidence.

When thorough formative evaluation is not possible, the only recourse is summative evaluation. Here statistical, rather than direct experimental investigation must be used to evaluate the efficacy of the procedures or treatment being developed. For example, the pharmaceutical industry is faced with developing new drugs with only limited guidance from formative evaluation—row 2 of the 3 X 3 matrix. Accordingly, a research–filtered methodology featuring a thorough summative evaluation is the approach of choice for assessing pharmaceutical effectiveness. Although effectiveness information is obtained, often little is learned about precisely how a drug does or doesn’t work (Valenstein, 1998), leaving the conclusions described in cell 2C of the 3 X 3 matrix. Accordingly, when assessing instructional programs, exclusively relying on FDA–like summative evaluation protocols may not necessarily be the most informative approach.

A thorough research–guided formative evaluation applied to designing and building instructional programs tells us more than that the “mean” experimental child performs better than the “mean” control group child. It tells us *how* the program components work separately and together, and whether or not it is effective with each individual. By setting formative

evaluation criteria high, we may be able to ensure that nearly all children who use a program so developed succeed. This is quite a different statement than saying the experimental group performed significantly better (statistically) than did the control group. We are all aware, that with a large enough “N,” small absolute differences between groups can produce highly significant results.

We would not board an aircraft based on a design that demonstrated that, “on average,” time aloft for one group of aircraft was greater than for another “control group” of aircraft. Should we be satisfied using a reading or other instructional program that works only on average better than another instructional program, even if that outcome has been “scientifically” determined? We argue that it may be a better scientific – or social – goal to produce educational programs that are the product of a thorough formative evaluation, and therefore must “fly” with each individual learner, one learner at a time. The important contribution that can be made by rigorous formative evaluation to predicting program outcomes, especially for individual learners, should not be overlooked.

SOME ESSENTIAL FEATURES OF A SCIENTIFIC FORMATIVE EVALUATION

In a different paper (Twyman, et al, 2004) we described the scientific formative evaluation approach employed by Headsprout as applied to the development of a beginning reading program. We distinguished scientific formative evaluation (row 3 of the 3 X 3 matrix) from other types of formative evaluation (row 1 and 2 of the 3 X 3 matrix), as illustrated in the following edited excerpt:

One typical approach to instructional design is to apply a top-down process. The goal is identified, broken down into smaller steps, and checked for “social agreement” —

do experts, or at least those with some familiarity, think it's reasonable? [row 1 or possibly 2 of the 3 X 3 matrix]; the program is then written in its entirety and tested with students. At times, designers will take data, note if their students fail, and redo some portion of the program [row 2 of the 3 X 3 matrix]. This often occurs in the context of field-testing with groups of learners. If the overall group tends to meet the goal [often judged by consensus and learner emotional reaction], then the product is considered finished. Many designers and curriculum publishers fail to perform the last two steps of minimal test and revision [staying firmly in row 1 of the 3 X 3 matrix]. Such a program may purport to present content that is derived from scientific principles, but the program itself does not meet the criteria for a scientifically developed program. The program may even provide a better outcome than some alternative approach against which it is compared, but still, the program cannot be considered scientifically developed.

Markle and Tiemann (1967) described a different instructional programming process. They noted that the entire instructional design process determines whether an instructional product will fulfill its vision. Markle and Tiemann took the position that a rigorous scientific control-analysis system is necessary for successful instructional design. Nevertheless, that recommendation has seldom been followed (Cook, 1983; Markle, 1969; 1990). One reason may be that there are few examples of its application on a large scale that can serve as a guide to others who are interested in producing quality instructional materials. Another, perhaps greater, obstacle is the time and expertise required to fully implement all elements of a scientific instructional design process. Markle and Tiemann's program development process can be slightly updated and summarized as follows:

1. Perform a content analysis. Content is examined and classified as to the type of learning involved (e.g., strategy, principle, concept, verbal repertoire, sequence or algorithm, multiple discrimination, paired associate, kinesthetic repertoire, chain or motor response [after Tiemann & Markle, 1991]).

2. State the objectives. Clear, measurable objectives are developed that reflect the content analysis and the overall goal of the program.

3. Determine the criterion tests. Tests are constructed against a standard that often involves both accuracy and frequency criteria (see Lindsley, 1997). The tests are developed for each teaching activity or routine within a lesson segment, for each lesson, for blocks of lessons, and finally for the program.

4. Establish the required entry repertoire. Given what is to be learned, determine the skills needed to progress through the program. Entry repertoires are the specific prerequisites skills needed for success, not simply prerequisite experiences (such as taking a "prerequisite" course without actually acquiring the behaviors identified in the course).

5. Build the instructional sequence. The content analysis and the criterion tests are used as a guide to produce instruction that will result in learner behavior that meets specified criteria.

6. Use performance data to continually adjust the instructional sequence (5) until it meets the objectives (2), as measured by the criterion tests (3), which reflect the content analysis (1).

7. Build in maintaining consequences. Plan for the different types of motivation that will be required, both program extrinsic and intrinsic (after Goldiamond, 1974). Use performance data, including affective data, to test and revise the motivational variables.

The learner begins at 4 (Entry Repertoire), goes to 5 (Instructional Sequence), and is evaluated at 3 (Criterion Tests) to determine if 2 (Instructional Objectives) has been reached. As is evident, the student does not progress through a sequence in the same way as the program was built. Nor is the program written in its entirety before it is tested. In this

approach the learner's behavior shapes the program until nearly all learners meet the specified criteria.

That is, the performance data (6) are used to continually adjust the program until nearly all learners meet the mastery criteria. To continue:

...All elements of the program are tested for effectiveness, and if the criteria are not met, alternative strategies are built and tested. The process iterates until all criteria are met, with performance always measured against a set of standards. The sequencing of program steps and their relation to the learner's behavior is explicitly identified, thereby generating new knowledge: about both the program and about learner behavior. This process continues and becomes aggregated as the "chunks" of the program units change in size (e.g., for Headsprout, a segment of a lesson, a lesson, groups of lessons, and the entire program). The research is based on individuals, and therefore can be generalized to individuals (Goldiamond & Thompson, 1967, reprinted 2004; Sidman, 1960).

...It is the job of the learning scientists and instructional designers to consider these factors when designing instruction, testing program segments, revising sequences, and interpreting learner data. It is important, therefore, to carefully control and vary how stimuli are presented, their sequence of presentation, the salience of the stimuli, the learner's history of responding to the alternatives, how the response request to the learner is made, and the consequences of responding to all alternatives (see Markle, 1990; Ray, 1969; Ray & Sidman, 1970; Stikeleather & Sidman, 1990).

This careful control analysis characterized by a scientific formative evaluation as represented by row 3 in the 3 X 3 matrix not only provides the predictive confidence educators need to make important decisions regarding the curriculum they select, but may also provide answers as to why a particular curriculum is effective. For example, in the online beginning reading program Headsprout Early Reading™, a single learner makes about

200 meaningful responses per 20 minute lesson, about 10 per minute. That means approximately 16,000 responses are individually collected and analyzed during the course of the program. As of this writing, across all learners, well over 100 million instructional interactions or "learn units" (after Greer, 2002) have been collected and regularly examined in an effort to understand how learners interact with the program and to continually improve it. This latter point should not be overlooked. For within the context of a thorough formative evaluation, experimental control and analysis of learner behavior can lead to new behavioral insights. Indeed, work from Headsprout's laboratory has experimentally replicated and extended procedures that can ensure successful discovery learning in the context of online educational programs (Layng, Twyman, & Stikeleather, 2004b).

LARGER SCALE FORMATIVE EVALUATION

There is yet another way a more scientific formative evaluation can play an important role in improving educational productivity, that is, on the district and school level. Though rare, some school districts (see for example Johnson, 2003) are working diligently to put in place sophisticated data-gathering instruments and frequent assessment so that educational practices at the classroom level can be tested, revised, and tested again until the practices are successful as measured against a standard. As a result, entire school districts have begun to make progress in closing the achievement gap between majority and minority students (Mozingo, 2003; Muffet et al, 2003), often with both achieving at much higher levels when careful formative evaluation practices have been used over time. (See also Farris, Carnine, and Silbert's (1993) description of formative evaluation being applied by the

Mattawan, Michigan school district to produce considerable district wide progress and a very high level of student achievement.)

CONCLUSION

Today, examples of commercial instructional programs built through a scientific process of thorough formative evaluation are rare (but, see Layng, et al, 2003, 2004a; Twyman, et al, 2004). Currently, a more common practice is to design and write materials that correspond in some identifiable way with previous research or authoritative consensus. Once written, the materials are perhaps tested with simple comparative or pretest vs. posttest studies in an attempt to provide at least some evidence that the materials may be effective, cell 2B in the 3 X 3 matrix. Consequently, as the movement toward more scientific practices continues, and given that most current reading and other instructional programs have not been scientifically developed, thorough summative evaluation, column C in the 3 X 3 matrix, will remain as the scientific method of choice for determining relative success for most programs. However, it should not go unappreciated that a rigorous control–analysis formative evaluation, row 3 in the 3 X 3 matrix, offers an equally scientific and arguably more reliable method of not only measuring, but of ensuring, program success.

Indeed, in an important recognition of the importance of a thorough formative evaluation, Wilbur Wright, when describing the use of the first wind tunnel to test wing designs, questioned whether powered flight would ever have been achieved without it:

It is difficult to underestimate the value of that very laborious work we did over that homemade wind tunnel. It was, in fact, the first wind tunnel in which small models of wings were tested and their lifting properties accurately noted. From all the data that Orville and I accumulated into tables, an accurate and reliable wing could finally be built. ...In fact,

the accurate wind tunnel data we developed was so important, it is doubtful if anyone would have ever developed a flyable wing without first developing this data...

In any case, as famous as we became for our "Flyer" and its system of control, it all would never have happened if we had not developed our own wind tunnel and derived our own correct aerodynamic data.*

Similarly, instructional programs whose designs have evolved from “educational wind tunnels” may provide not only a new “gold standard” for the application of scientific methods in the design and evaluation of instructional practices, but products not possible without it. Accordingly, scientific formative evaluation not only deserves the support of all those interested in the intersection of rigorous science with educational practices, but perhaps offers an even surer route to leaving no child behind.

* <http://www.wrightflyer.org/WindTunnel/testing1.html>

REFERENCES

- Bernard, C. (1865). *An Introduction to the Study of Experimental Medicine*. Translated by Greene HC. New York: Macmillan [Reprinted in 1927].
- Bloom, B.S., Hastings, J.T. and Madhaus, G.E. (1971). *Handbook on the Formative and Summative Evaluation of Student Learning*. New York: McGraw–Hill.
- Cook, D. A. (1983). CBT’s feet of clay: Questioning the informational transmission model. *Data Training*, 3 (12), 12-17.
- Coalition for Evidence–Based Policy (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Retrieved March 14, 2004, from <http://www.ed.gov/rschstat/research/pubs/rigoroussevid/index.html>
- Farris, H., Carnine, D., Silbert, J. (1993). Learning is our business. *The American School Board Journal*, December, pp 31–33.
- Goldiamond, I. & Thompson, D. (1967, updated 2004). *The blue books: Goldiamond & Thompson’s the functional analysis of behavior*. P. T. Andronis (Ed.) Cambridge, MA: Cambridge Center for Behavioral Studies.
- Greer, R. D. (2002). *Designing teaching strategies: An applied behavior analysis systems approach*. San Diego, CA: Academic Press.
- Habicht, J. P., Victora, C. G., & Vaughan, J. P. (1999). Evaluation designs for adequacy, plausibility, and probability of public health programme performance and impact. *International Journal of Epidemiology*, 28, 10–18.
- Johnson, A. (2003). Building the information foundation. Presented at the Council on Great City Schools, 47th Annual Fall Conference, Chicago, IL, October 22–26.
- Johnston, J. M. & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.) Hillsdale, NJ: Lawrence Earlbaum Associates.
- Layng, T. V. J. (1995). Causation and complexity: Old lessons, new crusades. *Journal of Behavior Therapy & Experimental Psychiatry* Vol. 26, No. 3, pp. 249–258.
- Layng, T. V. J., Twyman, J. S., & Stikeleather, G. (2003). Headsprout Early Reading: Reliably teaching children to read. *Behavioral Technology Today*, 3, 7–20.
- Layng, T. V. J., Twyman, J. S., & Stikeleather, G. (2004a). Selected for success: How *Headsprout Reading Basics™* teaches children to read. In D. J. Moran and R. W. Malott (Eds.) *Evidence based education methods*, St. Louis, MO: Elsevier/Academic Press.
- Layng, T. V. J., Twyman, J. S., & Stikeleather, G. (2004b). Engineering discovery learning: The contingency adduction of some precursors of textual responding in a beginning reading program. *The Analysis of Verbal Behavior*, 20, 99–109.
- Lindsley, O. R. (1997). Precise instructional design: Guidelines from Precision Teaching. In C. R. Dills & A. J. Romiszowski (Eds.), *Instructional development paradigms* (pp. 537-554). Englewood Cliffs, NJ: Educational Technology Publications.
- Markle, S. M. (1967). Empirical testing of programs. In P. C. Lange (Ed.), *Programmed instruction: Sixty–sixth yearbook of the National Society for the Study of Education: 2* (pp. 104–138). Chicago: University of Chicago Press.

- Markle, S. M. (1969). *Good frames and bad: A grammar of frame writing*. (2nd ed.). New York: Wiley.
- Markle, S. M. (1990). *Designs for instructional designers*. Champaign, IL: Stipes.
- Markle, S. M. & Tiemann, P. W. (1967). *Programming is a process: Slide/tape interactive program*. Chicago: University of Illinois at Chicago.
- Mozingo, T. (2003). Project acceleration: The advanced academic story. Presented at the Council on Great City Schools, 47th Annual Fall Conference, Chicago, IL, October 22–26.
- Muffet, G. & Wimberley, L. (2003). Decreasing the achievement gap: What works now? Presented at the Council on Great City Schools, 47th Annual Fall Conference, Chicago, IL, October 22–26.
- National Institute of Child Health and Human Development (2000). Report of the National Reading Panel. *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction: Reports of the Subgroups* (NIH Publication No. 00–4754). Washington, DC: U.S. Government Printing Office.
- Neuman, S. B. & McCormick, S. (1995). *Single-subject experimental research: Applications for literacy*. Newark, DE: International Reading Association.
- Neuman, S. B. & McCormick, S. (2002). A case for single subject experiments in literacy research. In M. L. Kamil, P. B. Mosenthal, P. D. Pearson, & R. Barr (Eds.) *Methods of literacy research* (pp. 105–118). Mahwah, NJ: Lawrence Erlbaum Associates.
- No Child Left Behind Act of 2001, Pub. L. No 107–110 Section 1201, 115 Stat. 1425 (2002). Retrieved March 14, 2004, from <http://www.ed.gov/policy/elsec/leg/esea02/pg4.html>.
- Paik, S. J. (2004). Evidence-based reform: Experimental and quasi-experimental research considered. Paper presented at the National Conference on *The Scientific Basis of Educational Productivity*, sponsored by the American Psychological Association and the Mid-Atlantic Regional Education Laboratory, the Laboratory for Student Success (LSS), May 13–14, Arlington, VA.
- Ray, B. A. (1969). Selective attention: The effects of combining stimuli which control incompatible behavior. *Journal of the Experimental Analysis of Behavior*, **12**, 539-550.
- Ray, B. A., & Sidman, M. (1970). Reinforcement schedules and stimulus control. In W. N. Schoenfeld (Ed.), *The theory of reinforcement schedules* (pp. 187–214). New York: Appleton-Century-Crofts.
- Scriven, M. (1974). Evaluation perspectives and procedures. In James W. Popham (Ed.), *Evaluation in education: Current application*. Berkeley, CA: McCutchan Publishing Co.
- Simmons, D. C., & Kame'enui, E. J. (2003). *A consumer's guide to evaluating a core reading program grades k–3: A critical elements analysis*. Institute for the Development of Institutional Development, College of Education, University of Oregon, Eugene: OR.
- Sidman, M. (1960). *Tactics of scientific research: Evaluating experimental data in psychology*. Boston, MA: Authors Cooperative, Inc.
- Sidman, M. & Stoddard, L. T. (1966). Programming perception and learning for retarded children. In N. R. Ellis (Ed.) *International review of research in mental retardation, Vol 2*. NY: Academic Press, pp. 151–208.

- Stikeleather, G. & Sidman, M. (1990). An instance of spurious equivalence relations. *The Analysis of Verbal Behavior*, **8**, 1–12.
- Tiemann, P. W., & Markle, S. M. (1991). *Analyzing instructional content*. Champaign, IL: Stipes.
- Twyman, J., Layng, T.V.J., Stikeleather, G. and Hobbins, K.A. (2004). A Non-linear approach to curriculum design: The role of behavior analysis in building an effective reading program. In W. L. Heward et al. (Eds.), *Focus on behavior analysis in education*, Vol. 3. Upper Saddle River, NJ: Merrill/Prentice Hall.
- U.S. Department of Education (n.d.) *Proven methods*. Retrieved March 14, 2004, from <http://www.ed.gov/nclb/methods/index.html?src=ov>
- Valenstein, E. S. (1998). *Blaming the brain: The real truth about drugs and mental health*. NY: The Free Press.

Table 1. A 3 X 2 Matrix Depicting Three Corresponding Levels of Verification for Formative and Summative Evaluation as They Relate to Claims of Effectiveness

	Formative Evaluation: Basis for Program Revision	Summative Evaluation: Basis for Outcomes Assessment
<p>Experiential</p> <p>Derived from philosophy or personal experience</p>	<p>Consensus of best practices, experience, point of view. Little or no testing during developmental process itself.</p> <p>Design revisions based on consistency of content with prevailing point of view.</p> <p>May employ limited tryouts that result in some program revisions. Clarity of communication typically the issue.</p>	<p>Correspondence to a point of view – philosophy or personal experience.</p> <p>Evaluation based on anecdotal evidence, look & feel, personal satisfaction, testimonials</p>
<p>Evidence Based</p> <p>Derived from scientific principles or Comparative measures</p>	<p>Consensus of best practices, experience, point of view, but design largely based on previous research, which may come from a variety of disciplines; may be on elements found in program content and not program itself.</p> <p>Design revisions often based on consistency of content with prevailing point of view; may employ checks for adherence to research.</p> <p>May employ limited tryouts that result in some program revisions. Clarity of communication typically the issue.</p>	<p>Pretest vs. posttest measures, meta-analysis, simple comparison studies –not employing random assignment or other controls</p>
<p>Scientific</p> <p>Application of scientific methods</p>	<p>Consensus of best practices, experience, point of view; design may or may not be initially based on previous research; may come from a variety of disciplines. All elements of program tested for effectiveness; if fails criteria, alternative built and tested; process iterates until criteria met. Performance is always measured against a set of criteria. Sequence of program steps and the relation of behavior to the sequence is explicitly identified, thereby, generating new knowledge. Process continues and is aggregated as the "chunks" of the program units change in size (e.g. a segment of a lesson, a lesson, groups of lessons, the program). Research based on, and systematically replicated with, individuals; thereby, can generalize to individuals (Nueman & McCormick, 2002; Sidman, 1960).</p>	<p>Randomized controlled group studies, measured against other programs, standard or placebo.</p> <p>(See Paik, this volume, for a detailed discussion of outcomes assessment.)</p>

Table 2. A 3 X 3 Matrix Describing the Relation between Formative & Summative Evaluation in Program Design & Outcomes Assessment

Approaches to Summative Evaluation: Basis for Outcomes Assessment				
	A. Experiential – Assessment	B. Evidence Based – Assessment	C. Scientific – Controlled Group Research & Assessment	
Approaches to Formative Evaluation: Basis for Program Revision	1. Experiential – Program Development	<p>Cannot predict group or individual performance.</p> <p>Works or not with groups or individuals purely subjective; a matter of opinion; argued on point of view –a matter of social agreement</p>	<p>Provides some indication that the program may be effective with a group; but</p> <p>Cannot confidently predict group or individual performance.</p>	<p>Can confidently predict group performance; but</p> <p>Cannot predict individual’s performance (Sidman, 1960).</p> <p>If works or not, not clear what program elements, alone or together are responsible.</p>
	2. Evidence Based – Program Development	<p>If limited tryouts, may indicate that the program might work with those tested; but</p> <p>Cannot confidently predict group or individual performance.</p> <p>Still primarily a matter of social agreement, but has some validity by relation to past research and perhaps limited tryouts.</p>	<p>Provides some indication that the program may be effective with a group; but</p> <p>Cannot confidently predict group or individual performance.</p> <p>If works, not really clear why, if it does not work, can lead to re-evaluation of principles or the way they were applied. Not clear where the problem is.</p>	<p>Can confidently predict group performance;</p> <p>Cannot confidently predict individual’s performance.</p> <p>If works or not, not clear what program elements, alone or together are responsible, but can lead to re-consideration of principles or the way they were applied.</p>
	3. Scientific – Controlled Individual Research & Program Development	<p>Able to predict group performance based on individual performance; and</p> <p>Can confidently predict individual’s performance.</p> <p>Since program able to predict individual’s performance, some prediction of group performance implied; may have some validity by relation to past research.</p>	<p>Able to predict group performance based on individual performance; and</p> <p>Can confidently predict individual’s performance.</p> <p>If doesn’t work, issues are in transfer, —able to identify and isolate variables to change and revise for retest. Individual data not lost and can be analyzed in relation to outcome.</p>	<p>Can confidently predict group performance; and</p> <p>Can confidently predict individual’s performance.</p> <p>If doesn’t work, issues are in differences in formative criteria & summative measurement instruments, —able to identify and isolate variables to change and revise criteria & program for retest, or to revise summative measurement instruments. Individual data not lost and can be analyzed in relation to outcome.</p>

The level of verification for each type of evaluation is indicated by the numbers 1–3 for formative evaluation, with row 3 representing the most thorough; the letters A–C indicate the level of verification for each type of summative evaluation, with column C representing the most thorough. Cell 3C represents the intersection of greatest verification for formative & summative evaluation.