The Struggle to Reform Education: Exploring the Limits of Policy Metaphors

CSE Technical Report 576

Eva L. Baker
CRESST / University of California, Los Angeles

August 2002

Center for the Study of Evaluation
National Center for Research on Evaluation,
Standards, and Student Testing
Graduate School of Education & Information Studies
University of California, Los Angeles
Los Angeles, CA 90095-1522
(310) 206-1532

Project 2.5 Quality Education Forum
Eva L. Baker, Project Director, CRESST/ University of California, Los Angeles

The work reported herein was supported under the Educational Research and Development Centers Program, PR/Award Number R305B60002, as administered by the Office of Educational Research and Improvement, U.S. Department of Education.

THE STRUGGLE TO REFORM EDUCATION: EXPLORING THE LIMITS OF POLICY METAPHORS[1]

Eva L. Baker
CRESST/University of California, Los Angeles

Educational Policy and Metaphorical Interpretations

Policy initiatives are almost always conveyed in shorthand, in a form of metaphor, so that the public can understand without difficulty what is proposed, so as to enable proponents to develop political support. To paraphrase Aaron Wildavsky (1979), in policy formation, the trick is to create the optimal projective test, one that allows agreement to vague precepts so that the bargaining process of politics does not get bogged down in details of program. This strategy has become even more important as the process of policy development and the power of the information media are increasingly entwined. Public opinion is a dependent measure in an ongoing policy time-series design. Gauged to reflect successively refined messages, the process supports messages that are short, staccato, and memorable. In this paper, I wish to consider the metaphors guiding educational policy now and their various interpretations. Then I will focus on what needs to be done if these educational metaphors are to be more than sound bites. I intend to focus on three closely linked metaphors to make my points. The general question on the table is whether it is best to clarify interpretations of these metaphors and transform them into operational tasks or whether we are better off leaving things vague. The answer will influence the probabilities around the 2010 educational futures.

Metaphor 1: All Children Can Learn

Of course all children can learn, have learned, and will learn. We need to know what this metaphor means in educational policy and how it may be interpreted in practice. I contend that there are better and worse interpretations and that their implications for educational prospects vary enormously.

As we parse the statement "All children can learn," first focus on the "all." This part of the statement is intended to raise expectations that every child will be

[1] A revised version of this report is scheduled to appear in H. Everson (Ed.), *Education in 2010.* New York: College Board.

effectively served by schools—by implication, a new state of affairs. In public rhetoric, a corollary has been offered that no individual child will be "left behind." Although evocative, perhaps a more apt reading of the line is that no *identifiable group* will be systematically served less well than any other. Such a reading provides clear policy guidance for the distribution of resources. It implies that schools will level the playing field for children who come in with disadvantages in skills and prior knowledge. It also implies that the educational system will distribute its expectations fairly and that high standards are desired for every child.

How else has the "All children" statement been interpreted? Some take it to mean that every child's learning will reach the same level of measured achievement; that is, no child will be left behind in the achievement of outcomes.

Let me cite an example of discourse intended to clarify the issue. At a recent CRESST conference (1999), one researcher, Daniel Koretz of RAND Corporation, was describing the interpretation of achievement data in light of the issue of examinee variation (Koretz, 1999). A polarized and passionate discussion between Dan Koretz and the audience ensued on the topic of variation in scores. Aside from the variable effectiveness of teachers, Koretz suggested that even in the most effective programs, there still would be individual differences in the achievement attained by students. More than one audience member held that such variation was unacceptable. "All children" meant every child needed to obtain the same level of expertise. It was also inferred that Koretz was talking about group rather than individual differences. This misunderstanding is similar to the uproar in the public as it rediscovers that not all children are reading at grade level and decides that a goal for schools must be that all children should read at the 80th percentile level. These goals express a desire for higher performance, but when they are made explicit, they inadvertently set impossible objectives for schools. Unfortunately, the result of failing to meet objectives is well known—loss of confidence and a publicized search for alternatives. Transformations of this metaphor into unachievable goals have real and negative consequences.

Let us return to the discussion of the goal of relatively no variance around a criterion. Low variance can be attained on a set of particular, very demanding tasks when training is very effective. Training (instruction) is successively revised until it achieves the desired outcomes. Multiple cycles may be required, and time must be permitted to vary. Another technique to reach high standards for all students is to eliminate any trainees who do not meet interim goals. They are washed out of such

systems, as in Navy SEAL training, where it is reported with great pride that only a very few, say "5%," reach criterion. The remaining candidates not only are left behind, but also are not allowed to finish the race, a result not dissimilar to some of the tracking programs used in European and Asian educational systems. It is one sure way to handle variation. One way, then, to achieve "all children" would be the following: (a) There are highly trained instructors; (b) learners are strongly motivated to stay in the "system"; (c) pre-instruction selection has been rigorous; and (d) measures of attainment are clear and unambiguous. These attributes are not part of most examples of the American educational system.

Another approach used to reduce variation when educating students in high-performance domains is to emphasize selection. Students apply to an educational program, such as law school, and a small subset is invited to enroll. The program is designed to support student success, even if time varies. "Washing out" becomes an economic disincentive. Consider medical education as an example of this category. Instruction is neither clear, short-term, nor targeted, as in the Navy SEAL example above, but two factors enable the system to achieve high performance. First, the motivation and background of students compensate for instructional inadequacies. Second, students exercise some serious choice in their selection of specializations, a procedure that allows for the optimization of requirements, individual talents, and preferences. Again, this approach does not correspond to the American educational model.

A third example of achieving low variation around a high criterion occurs when the outcomes vary for individuals. Think about a developmental preschool, where goals for some 4-year-olds involve interaction, for some pre-reading, and for others the decoding of printed sentences. The variation is handled exclusively through the definition of goals and measures. Yet all children demonstrate their learning of what they were focused on. How well does this approach map to usual school expectations?

The point of juxtaposing these strategies against typical approaches used in schools is to suggest that the literal interpretation of "High Standards for All Children" cannot be achieved in anything like the current system. Similarly, it means that if we don't clarify the metaphor, we are sure to fail.

How about a simple clarification? What if "all children" means that we intended to reduce the predictive value of race, social class, and other background

for school achievement variables? This is an interpretation that gets to the heart of the metaphor and can be achieved.

Metaphor 2: High Standards, Aligned Systems

The "All children" metaphor is located in another set of metaphors: high standards and aligned systems. High standards, sometimes amended by the qualifier "world-class," describe the outcomes of the system. Standards is a synonym for terms such as goals and objectives, used in prior policy mandates. Standards describe what is to be taught and learned and how well students perform. They also imply how attainment will be documented. The term "aligned system" describes the operation of a standards-based system—a coordinated effort to fit together key components to meet desired goals. In some ways, the idea of "aligned" is redundant with the term "system," where components are intended to coordinate. In education, we have been invoking the term "systems" more as a desire than a reality.

The "standards" metaphor is the launching point for the identification of goals (that is, general content) to be achieved in schools. One question involves *who* should do the "spelling out." In the late 1980s and early 1990s, the identification of high standards was thought best to be accomplished by content specialists, people who understood the subject matter and could discern what constituted high quality. To avoid the specter of federal intervention, the standards developed by such groups—for instance, the American Association for the Advancement of Science (AAAS) or the National Council of Teachers of Mathematics (NCTM)—were to serve as a resource for (or were adopted wholesale by) states and school districts. How did the process work? First, levels of specificity varied among the standards and among subject matters. Some standards described very general goals, such as writing clear, persuasive text; at another level, a standard might describe, in some detail, the historical perspective that was intended to be taken in a particular analysis. A second set of problems emerged because each subject matter group was developing its own set of standards and looking inward. As a result, most subject matter areas generated far too many standards on too many topics to be feasibly taught. Schools didn't have the number of instructional hours to meet standards for one subject matter area, let alone four or five. Moreover, no procedure, algorithm or guidance was in place to permit coordination among subject areas. So choices were made in odd ways. Politics and metaphor merged. Standards for reading and mathematics were interesting cases, for political interpretations (back-to-basics) were partially

supported by visible research. Furthermore, there were numerous instances of academic professionals pushing the envelope of credible content. “Why should we teach algebra or calculus,” some questioned, “when in the true spirit of constructivism, students should reinvent algebra if it is warranted by applied problem-solving tasks?” Such viewpoints, when paired with low test scores, rapidly raised public support for the basics.

The connection between high standards and cognitive psychology also yielded less-than-desirable outcomes. In part, cognitive language was used without clear examples that illustrated how standards were being raised. The uses of cognitively referenced terms such as knowledge acquisition, integration, problem identification, teamwork, and learning to learn became, for many school boards, soft signals for laxity (that is, lower rather than higher standards), even though the professionals meant just the opposite, and even though problem solving, teamwork, and self-management skills, along with basic skills, were reported by employers to be urgently needed in the workplace. The cognitive demands that were thought to operationalize the metaphor were rejected. The rejection occurred both because of patent miscalculation by the educational experts and because compelling examples could not be provided to illustrate what was really meant. Ironically, higher standards were rejected because they were not well understood.

Part of the idea of high standards was a corollary to content goals, something called “performance standards.” In the educational world, this term *had* meant either the criterion level or cut score necessary to be in a category (such as “pass” on a test) or otherwise deemed successful in a training program, *or* the set of criteria used to define such levels. A new interpretation was that a performance standard was an example of a desirable way of interpreting the content standards. These examples would define the boundaries of the content standards and provide guidance for assessment construction. If a student could read selected primary source historical materials and create an argument of a certain style, the standard of attaining a “general understanding of important themes of the Reconstruction period” would be explicated.

Performance standard developers soon learned that directly addressing even a few content standards resulted in a vast number of performance standards, far too many to implement. So, rather than reduce the number of standards, the plan became to combine multiple content standards into a single performance standard. The intention was that one assessment task would cover multiple standards. Though

motivated by practicality and trying to limit test development costs and the test-taking time required of students, combining standards into single assessment tasks created conceptual problems, the most important of which was that particular content standards could no longer serve as a clear guide to system development. Results from many standards were confounded, and thus it became impossible to report results in terms of the attainment of specific standards. The use of off-the-shelf tests as measures of standards similarly provided information that could not be neatly tracked back to the initiating standards. By combining many content standards into single performance standards and confounding assessment tasks, the core logic of standards-based reform was vitiated. In practice, this meant that performance standards, instruction, and assessments represented clumps of content standards, thus reducing the clarity of feedback intended to guide classroom practice, and resulting in a system much like the one that preceded it.

A closely related topic is the goal of alignment—the idea that we shouldn't adjust one element of the system without considering its impact on the remaining set of components. The most conservative view about alignment holds that system components should not contradict one another and undermine goals. A more optimistic vision is that there would be a systematic plan for the interrelationship and mutual impact of system elements. Thus, in a fully aligned system, or even a system that just worked reasonably well, we would look at the goals for states, districts, and schools and make them compatible; we would address the resources needed to achieve system goals; and we would use measures that clearly assessed the standards. An indicator of educational reform would be the increasing alignment or documented relationship among elements in the system. Aligned systems also imply the ability to act deliberately on results of reasonably valid measures of achievement. The degree to which standards were achieved would give another, more powerful indicator of system alignment.

That was the general idea, and the metaphor was humming and well oiled. When attempting to address how to get systems better aligned, a major implementation error occurred. Educators began to attempt to address, simultaneously, all parts of the system. The logic was that if the whole system needed to function together, then everything should be fixed at once. That a system ultimately needed to be coherently functioning obviously did not imply that effort should be initiated on every element. Rather, some notion of priority should have been guiding implementation. Elements requiring longer development should have

been targeted, and those should have been addressed first. Instead, generally there was no sense of orderly development. Teacher quality, instructional resources, social coherence, accountability, parental involvement—all were given equal weight. As a result, few elements had enough resources to be properly addressed, and partial or fitful implementation characterized the development of the system components.

The second flaw in the idea of alignment is that there was, and is, virtually no idea of how to determine whether or not you have it. In a systems context, the lack of a method to determine the criterion state is fatal and not just a little ironic. Yet, in education this is the case. Alignment is a desired characteristic, and as a result, it is asserted on the flimsiest of evidence. On instructional alignment, decisions will be made topically. For example, if the wording of a standard, a section of text, and an examination question all include the term "geometric figure," then the system may be deemed aligned. If standards have been written and teachers have been given copies of them, the system can be claimed as moving toward "alignment."

The part of aligned systems of most interest to me is the relationship of the system goals to the measures taken to determine whether the system is achieving its goals. It is closely related to the discussion about combining content standards into a reduced set of performance standards. Let me summarize my concerns.

1. The disintegration of standards-based reform occurred when standards were combined into fewer performance standards, or when it was decided that a test could address sets of standards without looking at the item design and distribution.

2. This led to the use of off-the-shelf tests to measure standards, with reporting based on national norms.

3. The investment in accountability (especially with the expectation of rapid improvement) focused attention less on the domains implied by the content standards and more on raising test scores, using test preparation and motivational techniques. Short-term accountability requirements, thought to make tests "count," actually work against the goals of high-performing systems. They push investment to those attributes that yield short-term payoff and may leave untouched real instruction in the desired domain.

Choices about investments in education are usually politically rather than technically driven, based in part on public perception and the power of various constituencies. But the metaphor of alignment, the lining up of key factors, suggests a deliberate "moving around" and adjusting of components. Instead, political

pressures resulted in investment in parts of the system that might not have been most in need.

Management by Results

Let me go on to the third metaphor, "Management by Results"—an old song played in other venues and in other keys in education. The idea of management by results in a standards-based system is based on six major principles of action: (a) You know what you want; (b) you measure it well; (c) you make reasonable choices based on the interpretation of results; (d) everyone knows the consequences of the system; (e) management accounts for key system variables; and (f) the outcomes improve.

The previous discussion should make it clear that the articulation of standards as a metaphor and their operational definition undermined the stated intention of standards-based reform; the measurement system does not reflect the intentions of the system—that is, all children can achieve high standards. In addition, the interpretation of results does not reflect specific standards. As a result, interpretations are global—"we're doing well," "we're doing poorly"—but no method to focus on particular remedies is clear. Consequences of the system may be known—whether schools are classified into a "give help" or "disestablish" category, teachers get rewards, or students get certificates or summer school—but without clarity about the most effective change strategies for improvement.

Three issues deserve additional attention. First, let us assume that the standards could be measured well. If that were the case, we would still need to know that the classification into categories (receiving rewards, receiving sanctions) was sufficiently accurate for individuals and groups. Rogosa (1999) has prepared analyses demonstrating that this assumption may not be so. If classifications are inaccurate, consequences are misapplied, and the credibility of the system suffers.

Even assuming that classifications are accurate, there is the question of remedies. Are the "remedies" under control of the managers of the system? Besides the limits on resources to address professional development, teacher quality, children who learn at different rates, and so on, there are public preferences that need to be considered. It may be that the best approach empirically to improve reading would be one that would not be acceptable to some teachers or some groups of parents. In the present climate, many decisions reflect uninformed preferences rather than technically supported judgments.

Finally, what are practical ways of focusing on the poorest performing students? We don't mean to close the achievement gap by dropping the top. Instead, we must expect faster progress by those near the bottom of the distribution if they are to "catch up" and not be left behind. These are the very students that the managers of educational systems least well address. They are more mobile (and relatively few districts and states have information systems to monitor students' movement from school to school). These students often have limited English competency. They frequently are assigned the least well-prepared teachers. Their classrooms have less technology used in less productive ways. For management by results to work, the focus of effort needs to be specifically directed to students who have had low performance in the past. Yet, in reviewing the options available for these students, there is little specific evidence that there are available teachers either with the capacity to help them catch up or who can deliver powerful instructional tools to assist them. Management by results implies that people, given poor performance, will know how to fix the problem. In our system, some may know, but specific educational knowledge needs to survive a political gauntlet before it is implemented.

Metaphors Revisited: The Future of Educational Reform

Metaphor is an implied comparison, and I am alleging that broad educational policy statements imply comparison to other spheres of activity—business, training, mechanical systems—and that educational contexts differ dramatically from these settings. Figuring out the conflicts and cross-purposes of different interpretations is inefficient; it is a second-guessing exercise that happens too late to help. As a result, educators interpret these metaphors in different ways. Transforming metaphors into real operations at the outset seems to be the only way they can be productive. If procedures are not agreed on at the outset and clearly described, then activities will be inadvertently undertaken that undermine, rather than support, the intended goals.

Unfortunately, I think that the prognosis for a metaphor-rife reform is poor. If we are left in the present state with mismatched measures, I would say that the best that we can expect to see is short-term growth on measures, attributed to motivation, familiarity, test preparation, and basic skills. Growth on high standards will not be documented because it will be largely unmeasured and because the capacity of the system hasn't changed quickly enough. In any case, progress will

flatten out after 3 years or so. The public will infer that high standards cannot be achieved and that continued investment in the public system is not justified.

For the Future

Are there steps that can be taken immediately to provide tests of the hypothesis that educational reform can make a difference? One suggestion is to create and promulgate strong standards to judge the quality of assessment and accountability systems. Such standards could attempt to guide practice to be more concrete and defensible. Here is a preliminary set of standards for consideration.

- Articulate feasible goals and purposes to be achieved by the system.
- Design and document valid, interpretable indicators for particular goals or purposes.
- Evaluate over time the accountability system's effects on all of its goals.
- Plan symmetrical accountability affecting institutions, individuals, policy, management, teachers, and students.
- Combine multiple measures of achievement so that action is implied by results.

In addition to a set of accountability standards that can be used for self-study or for public or media review, more effort on basic research is necessary to get a concrete sense of what is desired, what is learned, and how we should best measure outcomes. I agree that alignment is key and that we need a different way to look at it.

To address alignment specifically, I have advocated (Baker 1997, Baker & Niemi, 1998a, 1998b, 1999) investment in a detailed map of school learning. This map would represent the types and elements of content, topics, principles, and procedures expected to be learned. We would also need to model various constraints (for example, linguistic requirements, cognitive demands, time limits, and social support) that influence acquisition and performance. By using such a representation, operational relationships among different tests, different instructional materials, software, and other resources could be defined. Search engine technology and advances in information science provide encouraging signs that we could accomplish a trial of the usefulness of this approach in a reasonable period of time. Such an approach would allow educational artifacts to be

characterized by various descriptive fields: linguistic, content, cognitive, and task characteristics.

This project calls for significant basic research. What are search parameters? How do these domains get modeled? How are strengths of relationships found? How is human development addressed? If even a partial map were available, then hypotheses about the relationship between instruction and various measures of learning could be tested. Right now, we are using far too global descriptions of the components of educational systems.

Undoing Damage

We also need a massive and skillful public relations campaign that says that many tests being used for accountability should not be bearing that responsibility and that inferences drawn from such measures are misleading. If an investment in better measures is made, we can create tests with tight linkage to content domains, cognitive demands, and task performance. Criteria for such assessments would include

- sensitivity to instruction, meaning that good instruction registers on the measure;
- high-quality technical analyses;
- longitudinal measurement;
- appropriate purposes identified and others explicitly excluded; and
- interpretive evidence on the utility of the test for every purpose and for subgroups of the tested students.

The future and quality of education depend upon our ability, with clarity, to transform our wishes, our hopes, and our metaphors into practices and activities that demonstrably affect students, teachers, and parents. Policy initiatives should articulate our goals and how we will know we have met them. It is time to break away from the development of vague initiatives and provide a focused analysis about what we want to do and why we think it will work.

References

Baker, E. L. (1997, Autumn). Model-based performance assessment. *Theory Into Practice, 36*, 247-254.

Baker, E. L., & Niemi, D. M. (1998a). *Design and development of a comprehensive assessment system. Identification and pilot testing of performance assessments and validity studies development*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Niemi, D. M. (1998b). *Design and development of a comprehensive assessment system. Summary of the 1998 Secondary Department Chair Forums and the Performance Assessment Design Institutes*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L., & Niemi, D. M. (1999*). Design and development of a comprehensive assessment system. Delivery of performance assessment tasks for field testing.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Koretz, D. (1999, September). *Assessment based reform: Taking stock.* Paper presented at the 1999 CRESST conference, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California, Los Angeles.

Rogosa, D. R. (1999). *Accuracy of year-1, year-2 comparisons using individual percentile rank scores: Classical test theory calculations* (CSE Tech. Rep. No. 510). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Wildavsky, A. (1979). *Speaking truth to power: The art and craft of policy analyses*. Boston: Little, Brown and Company.