Aligning Curriculum, Standards, and Assessments: Fulfilling the Promise of School Reform

CSE Report 645

Eva L. Baker
National Center for Research on Evaluation,
Standards, and Student Testing (CRESST)
University of California, Los Angeles

December 2004

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 1.2: Systems Design and Improvement: Ideal and Practical Models for Accountability and Assessment, Strand 1
Project Director: Eva L. Baker, CRESST/UCLA

The work reported herein was supported in part under the Educational Research and Development Centers Program, PR/Award Number R305B960002, as administered by the Institute of Education Sciences, U.S. Department of Education, and in part by Educational Testing Service.

The findings and opinions expressed in this report do not reflect the positions or policies of the National Center for Education Research, the Institute of Education Sciences, the U.S. Department of Education, or Educational Testing Service.

ALIGNING CURRICULUM, STANDARDS, AND ASSESSMENTS: FULFILLING THE PROMISE OF SCHOOL REFORM[1]

Eva L. Baker

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)/UCLA

Introduction

Throughout the last 40 years, educational policymakers have designed educational interventions for the purpose of improving the learning of disadvantaged students and of children with special needs, beginning with the Elementary and Secondary Education Act of 1965 (ESEA). More than 20 years ago, with the publication of *A Nation at Risk* (National Commission on Excellence in Education, 1983), American educational policy widened its focus to include the learning and achievement of all children, in part as a reaction to U.S. students' mediocre performance rank in international achievement comparisons. In the current climate, educational improvement is no less important an end. The recent version of reform, standards-based education, grew out of several linked events: (a) the action of governors interested in their states' economic competitiveness in the United States (National Governors Association, 1991); (b) the report of a congressionally appointed national panel on standards and assessment (*Raising Standards for American Schools*, National Council on Education Standards and Testing, 1992); (c) the National Education Goals Panel's (1991a, 1991b, 1991c, 1992, 1993) recommendations on goals and reporting; and (d) successively enacted legislation, including America 2000 (1991), Goals 2000 (1994), and the Improving America's Schools Act of 1994. The most recent and expansive legislation, the renewal of the ESEA, articulated in the language and policies supporting the No Child Left Behind Act of 2001 (NCLB, 2002), emphasizes the importance of the measured achievement of all students. It has raised the consequences of test score results while at the same time requiring more grade levels to be tested and more detailed reporting on the performance of groups within schools. Although there are numerous other provisions, involving requirements for teacher quality, the use of evidence in making decisions, and so on, NCLB early became known for its emphasis on testing and accountability. As the dominant, legislated form of

[1] Forthcoming as: Baker, E. L. (in press). Aligning curriculum, standards, and assessments: Fulfilling the promise of school reform. In C. A. Dwyer (Ed.), *Measurement and research in the accountability era.* Mahwah, NJ: Erlbaum.

educational improvement, NCLB enjoins states to rely on student test results as the primary information source to assess progress and to guide the improvement of learning. In a framework that emphasizes accountability as the path to growth, NCLB archetypically demands a system where responsibility for outcomes is located and sanctions (or rewards) are assigned. For the purpose of this chapter and for brevity, let's call this amalgamated set of strategies Results-Based Reform (RBR).

How should RBR work? The rhetorically preferred sequence of RBR begins with the specification of standards, or educational goals, operationally states performance standards, not unlike behavioral objectives of my youth (Baker & Popham, 1973), applies interventions (instruction, teacher development, motivation, and/or other resources), and publicizes the results of tests or other measures of performance of relevant groups (typically, but not always, students). It is planned that the first set of findings triggers an iterative cycle of (wise) inferences and (effective) actions inspired by successive and useful reports of accurate results. In the articulated plan, this system works, focusing each year on those who might be left behind, until after a 12-year cycle, all children meet explicit levels of proficiency—that is, 1st graders need to be proficient in 1st-grade standards, 7th graders in 7th-grade standards, and so on. It isn't just the 12th-grade students who will reach the achievement nirvana. In reality, it is likely that this approach will end somewhat sooner than planned, perhaps when asymptote is reached, a new test is brought in, the rules change, or new policy goals are posited. On the surface, the logical framework of the reform is straightforward: To be most effective, there needs to be quality and coherence in the array of stated goals or standards—they should combine to enable the student to exemplify the compleat learner, or something close. Furthermore, there must be a known relationship among standards, benchmarks or targets, tests, information received by teachers, their interpretation, and the generation of new, improved approaches. For the system to work right, the iterations should occur within a school year, or better still, within a unit or academic segment, rather than among years. Thus, measurement of results, a glowing signal for need of improvement, could take place many times in a term and would occur as close to the relevant instruction as possible. Once-a-year, formal measurement offers lean hope of fixing the system, except by extrapolation—teachers may redesign their year-long course in hopes of remedying performance yet to be exhibited by a new cohort of students. The students who already have been left behind will have to catch up in the next course. How will these students, who have done poorly, make

faster progress? Evidence about retention in a grade level (another model based on a "whole year" unit of analysis) confirms that this approach, as a remedy for social promotion, has neither breadth nor depth of scientific base (Hauser, 2004).

Among the assumptions of this general system, or what is now called a theory of action (Baker & Linn, 2004), are that goals can be usefully described, and instructional options are available and can be carefully selected and applied to improve outcomes that relate operationally to the standards. Over the years, this model has made multiple appearances on the education policy stage, in modest and lavish costume and in different roles. The star components of RBR have been subject to different interpretations.

For example, among the variations have been the optimal degree of specificity of learning goals, the details of description (or not) of curriculum, the unit of analysis chosen for improvement, and the forms and frequency of measurements and assessments. In addition, the clarity of expectations, the audiences for reports, and the consequences of following the process and achieving ends have toggled back and forth. Some changes are simple, lexical preferences—for instance, standard, goal, objective—and may presage (should we have them) the verbal analogies on tests in years to come: *goals : standards* as *behavioral objectives : performance standards.*

Yet, the underlying steps are always the same, whether inferred from the rationalism of Aristotle or the more explicit procedures of the curriculum *rationale* articulated by Tyler (1949). The system has a logical set of requirements and both chronological and looped sequences. Figure out what should be taught, be prepared to teach it, help students learn it, measure their learning, and continue the cycle until desired improvement is met. As my grandfather often said at his persuasive heights, "So what could be bad?"

Continuing on a personal level, the general RBR approach early on made sense to me, despite my strong reservations about what and how much can or should be engineered in learning. Looking back on my own work, RBR has pervaded much of what I taught, how I tried to conduct my own teaching, and my choice of R&D strategies in instructional design, measurement, and teacher education. It had nothing at all to do with my private life or my parenting. As a graduate student, I saw the RBR logic made concrete, albeit imperfectly and in micro form, in teacher education (Popham & Baker, 1970) and in the procedures and products of programmed instruction (Holland & Skinner, 1961; Lumsdaine & Glaser, 1960;

Markle, 1967). More important, I saw it used to produce reliably important learning. Without a doubt, I am an ancient partisan. In the broader scheme of things, the enterprise of instructional systems design (ISD), as practiced in military and business training, was not incidentally built on the identical paradigm (Baker 1972, 1973), accounting, perhaps in part, for the almost uniform faith accorded RBR by leaders in business and industry. Even common, small R&D strategies, like pilot tests in research studies and the practice of formative evaluation, use the same syntax—focused empiricism: Conduct trial, review data, revise, and when warranted, expand. As we all can attest, RBR has moved rapidly across the rhetorical and policy horizons to its present ascendance, in tractor-beam with accountability.

At a more earthly level, the theory of action of RBR in an accountability framework requires an ever-greater number of particular steps. When broad-based reform is the goal (as opposed to, let's say, a functioning instructional program), the components of RBR consist of (a) the announced benefits or sanctions associated with accomplishment; (b) goals or standards; (c) intended beneficiaries; (d) the desired level of operational attainment equivalent to targets; (e) the rate of progress of these attainments, usually expressed in percentages of the population reaching certain levels of competence, or Adequate Yearly Progress in NCLB-speak; (f) the inputs; (g) the resources; (h) the operations; (i) the measures; and (j) the obtained results, reported as required. The results trigger another series of events: (a) External inferences are made from the data (good, improving, disastrous, weak at occasional points), followed by (b) the invoking of consequences associated with accountability, (c) retargeting goals for all or subsets of students, (d) redesign of the reform implementation, (e) reallocation of resources, and (f) fielding the revised effort, *ad infinitum*, or until the individuals or institutions meet the goal, or it disappears, or the accountability system is revised.

A key part of the theory underlying this set of events assumes that there is adequate knowledge of the components and alternative courses of action by participants in the system, the wherewithal to implement them, and an acceptance of the power of the incentives and sanctions attendant to results.

Alignment

The focus of this piece is alignment. Alignment is the ether in which float the component parts of RBR. The logic of actions, the accuracy of inferences, and at the core, any reason at all to believe that systematic action will achieve positive results

in an RBR framework depend on alignment. So what is the present state of the concept of alignment, who wants it, how is it variously conceived, how has it been measured, and in what ways may it be given added utility—or reconceived?

Much discussion of alignment occurs as a general proposition—a system is either aligned in whole or it is not. Yet, even a cursory analysis of the number of required pair-wise relationships of the list of components above suggests that alignment is a massive undertaking. Even for one subject matter at one grade level, true alignment, or the explicit relationships leading to the management of instruction and testing, is beyond the capacity of most school programs. If required to document the logical or empirical (let alone scientific) evidence of relationships, we find the number of separate links is in the thousands. If the relationships, or the links among components, are further multiplied by the number of improvement cycles, the number and backgrounds of students, and the range of different organizational contexts of classrooms, schools and regions, personnel capabilities, and resources, we are swamped by the overwhelmingly large number of relationships that are required for effectiveness and efficiency.

The Quandary

Without a semblance of alignment, nothing hangs together. Goals may or may not be exemplified in practice, children may or may not learn what is expected, and test scores could represent standards or miss their mark entirely. Inferences about results could rarely be tightly justified, and subsequent findings may not respond to deliberate actions by students and educators. In an unaligned system, audiences can be misled by reports and may attribute change in results, or the lack of it, to inappropriate sources, a not unknown error in causal inferences. Deficiencies in alignment result in ambiguity that may affect some or all parts of the system, like an incubating virus—dangerous but not that obvious. In accountability contexts, we choose ways to improve performance, invoke sanctions with real-world consequences, and, over time, desire to help more students reach a full range of goals. Without adequate alignment evidence, we are left with luck and magical thinking as our tools of choice to improve education.

If alignment probably cannot be taken literally, even at the level of planning, what is the midpoint between tightly-matched processes and hocus-pocus? Where are the critical places for alignment to operate? Are there complementary or alternative formulations that will attain the wanted outcomes—that complex

learning becomes a clear and unequivocal result of schooling and the attendant preparations of teachers and students?

Grasping Alignment

In the olden days, in an analytic presentation or article, it was common to start the exposition with a definition. This appeal to authority usually cited *Webster's Dictionary* or the *Oxford English Dictionary*. The practice has changed. The equivalent, in the electronic era, is to rely on the Internet. Its authority comes not from precision but from sheer numbers. So, in preparation for this discussion, I Googled *alignment*. The search engine spit out a list of 2,800,000 hits for the exact word *alignment*, so I decided not to vary form class, for instance, the verb *align*. I sampled, as you might imagine, the list, exploring what seemed to be common and where very different interpretations were exhibited. In the most literal interpretation of alignment, components are *lined* up, arranged in a straight line. Like the Rockettes or the University of Iowa marching band, either of which is a sight to behold. How is such alignment achieved? As a precocious child, I had the honor of assuring the alignment of the Hollywood High School marching drill team. Each young woman was required to arrange her shoulders so that they were at even in the horizontal with those of her colleague to her left. I monitored each line, to assure that every marcher, undistracted by pompoms or the cheers of the spectators at football games and the Hollywood Christmas parade, maintained her relative position with the girl on her left, whether walking on sidewalks, turf, or, on occasion, up a flight of stairs. Alignment was managed by making sure each link was positioned in space as planned. Yet, we have already disclaimed the likelihood of success of such an approach in the complexity of educational reform.

Discounting the lining up in alignment, and eschewing the acknowledgment of magical thinking as the only alternative, I believe that my Google search helped me understand a reasonable midpoint for thinking about alignment—the various metaphors that can be used to interpret the term. Here are some Googled (and non-Googled) examples.

Chiropractic is the metaphor here (Figure 1). This sweet dog, apparently aging, needs to have its spine aligned. Alignment of the spine does not mean the same thing at the cervical or lumbar regions. Thus, adjustments are made to reduce, on the one hand, pain and perhaps to provide, on the other hand, a sense of well-being. In this picture, the bed (or the framework) is intended to be central, and to prevent the dog from falling out of alignment.

Figure 1. Max (copyright 2004 DK Cavanaugh). Reprinted with permission of D.K. Cavanaugh.

In Figure 2, a very frequent example of alignment, we have more than one feature needing to be adjusted—all four wheels must work together, and because of the variations in tires, cars, weight distribution, and wear and tear, alignment means not only that the wheels will go in the direction one wants, but also that they will proceed in a balanced way (wheel balancing is usually a component of alignment). Thus, linked relationships and balance represent two criteria of alignment.

Figure 2. http://www.fly-ford.com/StepByStep-Front-Series.html. Reprinted with permission of Marlo's Frame & Alignment, http://www.fly-ford.com

In the next version of alignment (Figure 3), we extend the notion of balance to the extreme. But in order to achieve such a balance, the practitioner must be aligned not only in the corporeal sense, but in other ways as well. Achieving some of the more difficult positions or *asanas* occurs only when the practitioner is more deeply aligned—mind (intentions), body, and spirit (a calm, a centering) working together. This kind of alignment comes from inside out rather than from a surface analysis.

One of the most common uses of the term alignment, and one of the most ancient, comes from the study of the heavens (Figure 4). Both astronomy and astrology, documenting regularities, calculating relationships and drawing inferences about the cosmos, depend on understanding how relationships are formed, how long they last, and when they may be expected to occur again. In the astrological interpretation, alignments of certain types—which planets, moon, time, and place—portend different futures.

Figure 3. http://www.powerofyoga.com/. Reprinted with permission of Sherri Baptiste Freeman, Baptiste Power of Yoga.

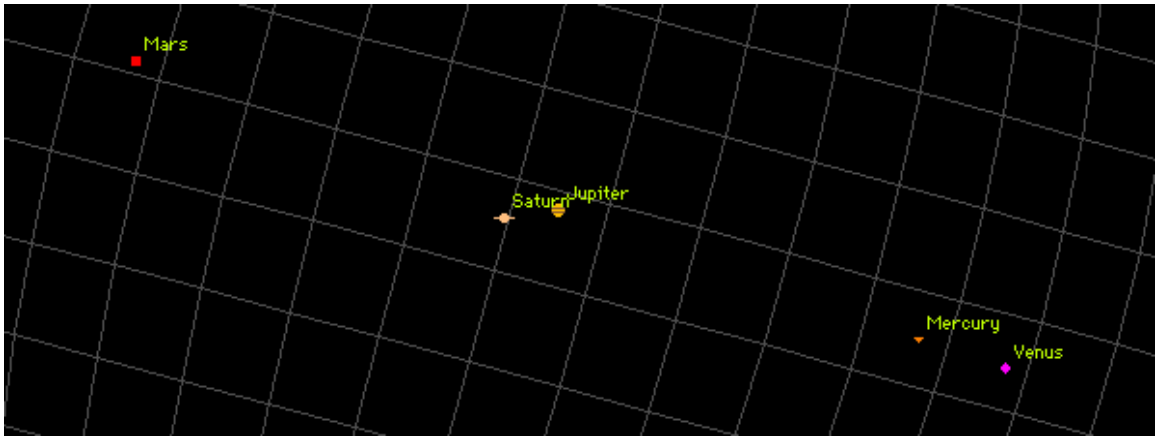

Figure 4. http://www.carinasoft.com. Created with Voyager II software. Reprinted with permission of Carina Software.

In the astrological version of alignment, one finds some common ground with educational practice—that is, the more general the prediction (you will feel angry at someone today, but your future is bright), the better. Thus, alignment in education has been claimed for instruction when it has been documented by survey that teachers have seen and read the content standards for which they are to be accountable,[2] or that one item of many performance standards is related to an overarching content standard, to wit, "The area of this square is _____" for a content standard that says "to demonstrate a deep understanding of geometric principles."[2] These examples illustrate that the general use of a term such as alignment, for instruction or testing purposes, encourages the most minimalist interpretations to be allowable. This minimalism developed from imprecision is exactly where policy has placed the world of practice. If, for instance, the law says a system will be aligned (and leaves alignment, as it should, to the state or local authority), and the state requires of its test vendor, for instance, tests aligned to standards, any evidence of the positive will count. The test will be aligned.

But the most important idea, a deeper truth, can be learned from this planetary type of alignment. It is that alignment is not forever. Planets fall in and out of alignment. The idea of "aligned system" then is only a temporal one and will change with changing emphases, resources, and, like the stars, time. The changes in educational alignment will be far less predictable, but change there will be. This suggests to me that the range of alignment (acceptable parameters) should be articulated, rather than imagining alignment to be, first, an on-off principle, and second, something stable, unchanging, and good for all time.

[2] References are omitted as a kindness.

There were many other examples of alignment, one of which was the DNA coils (Figure 5).

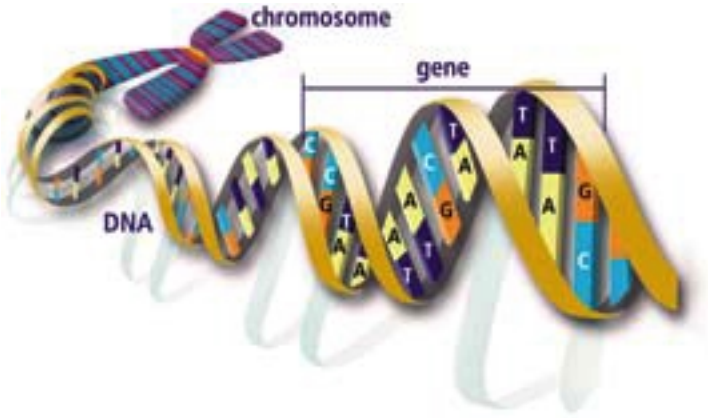

Figure 5. http://www.ornl.gov/hgmis. Reprinted with permission of the United States Department of Energy Human Genome Program.

Some of you know that at UCLA, we began a study of alignment that was modeled on this metaphor. Instead of the Genome, we called it the LEARNOME (Baker, Sawaki, & Stoker, 2002), and it derived from a paper given at ETS on the occasion of the Angoff lecture (November 1998). The idea was to find the smallest usable primitives that could characterize tests and instruction. The domains to be "mapped" involved content knowledge and skills, cognitive demands, linguistic requirements, tasks, and situational demands. There were other features that could be studied, for instance, the range of developmental trajectories through the space, and the way individual and institutional artifacts could be represented. In our preliminary work, really an illustration of how to take the linkages of alignment most literally, we were most advantaged by the work of linguists (e.g., Sawaki, Xi, Stoker, & Lord, 2004) and stymied by the complexity of the cognitive requirements. This experience, now in part translated into part of the 10-year plan for the Federation of Behavioral, Psychological, and Cognitive Sciences (Kelly, 2004), has led me to look at different approaches to alignment in educational reform as a set of metaphors that might be expanded, discarded, or adopted.

Metaphors for Alignment

Alignment as Congruence

Congruence is the easiest state of alignment to describe and the hardest to achieve acceptably. It means that each and every goal in the system is clearly specified and is measured completely and without added irrelevance. It means that instruction matches both the goals and the measures, and covers their intentions completely. It implies that deliberate action (i.e., instruction) explains a large proportion of increased performance. Congruent states of alignment can be achieved in highly focused systems. For example, as part of the qualification to be certified as a Navy SEAL, one of the elite forces in the military, the candidate must swim underwater for 50 meters on one breath. The goal is clear, the criterion or performance standard is set, the instruction involves practicing the behavior, with the subcomponents of controlling anxiety, keeping oxygen consumption low, and swimming at a reasonable rate. Congruence of the goal, instruction, and measurement is achieved. Note that there are metacognitive and emotional skills, as well as psychomotor skills, involved. Even here, congruence of alignment is difficult because of the important issue of transfer; that is, under what other conditions, physical, geographical, and emotional, will the SEAL need to be able to perform this task?

Congruence is especially difficult in two common cases in education: (a) where the goals are broad and generally stated, and (b) where there are too many goals to be adequately taught and measured. In those cases, the logical course of action is to focus on something other than the goals or standards as the basis for alignment. In practice, because the test or set of measures is seen as an operational definition of the domain of interest rather than a sample of it, the test is used as the guide, somewhat like the shoulder of the girl to the left in a drill team.

At best, this approach simulates system congruence by matching instruction to a subset of what is intended—the specific measures rather than the goals. Students are then given practice on the form and specific content of the tests used for accountability. Trivial modifications in wording of the examination are included in test practice materials, for instance, performing calculations on sets of oranges instead of apples. Pejoratively called teaching to the test, or teaching the test, this approach can be very effective in raising the scores on the particular measures in use. It is eminently rational, particularly if there are no clearer guidelines. It is true

that alignment of a sort is attained. I first thought that if the claim was made that students were proficient at the standards, then the price was inappropriate inferences about student performance and institutional quality. After more thought, it seemed that the price might be small if the public persists in believing that raised test performance is in fact the goal of education, and that the standards are merely documents for communicating in common language.

The real cost, however, is the learning that students are supposed to apply outside of school: transfer or application to new settings and conditions, with varied pieces of knowledge, help, time pressures, and criteria. It is possible, by examining tests of the same domain, or transfer tasks, to assess in part the degree to which any identified test is adequately substituting for the standards or goals, and thereby actually supports transfer (Koretz, 2003a, 2003b). One expects that students' performance will be lower on tests or measures differing in format or content examples, unless the measured test requirements include the content addressed by the alternative tests. For example, if students were expected on the official examination to solve equations with more than one unknown and the alternative examination included conceptually easier items (i.e., one unknown only), then it is likely that performance would be comparable. But for instruction to approach anything like efficiency, the number of standards prescribed and the clarity with which they are described must be limited, both so that instruction can be focused on them and so that the standards may be adequately measured (Commission on Instructionally Supportive Assessment, 2001). Congruence is partially inferred from results, but we will allude to the knotty problem of measuring indicators of alignment in a later section. Note that accountability systems with short intervals before imposed consequences assume a degree of congruence rarely realized in complex systems. So if congruence were adopted as the model of alignment what would be the summary risks and mitigations?

Risks of Congruence

- Unknown level of real learning
- Limits on domain learning and transfer
- Ethical questions
- Boredom and disengagement of students

Mitigations

- Systematic, breadth-and-depth sampling strategies
- Rotating constructs
- Publication of test items
- Evidence of rising test scores, thus political protection

New Metaphors

To stimulate thinking, and in no way to present fully worked out options, consider some conceptions or metaphors for alignment that may provide a usable approach to addressing the relationships among educational processes, results, and inferences.

Alignment as a Set of Correspondences

Correspondence is the state of being ". . . in agreement, harmony . . ." (Morris, 1981, p. 299). Correspondence permits analogies, functional rather than literal agreement, and a table of equivalences. For example, in the area of reading comprehension, it is possible to describe and illustrate the types of text and sorts of answers an individual is expected to provide to meet different levels of a standard, to make progress, and so on. One excellent example is the recently published volume from Equipped for the Future (EFF), *EFF Standards* (EFF Assessment Consortium, 2004), where standards, performance levels, and assessment tasks are admirably related and illustrated, essentially at the level of what goes with what. The benefit of a correspondence model of alignment is that it is at heart based on illustrations. The limitation of this approach, of course, is that it requires some sophisticated extrapolation about other examples that would serve the same function for common standards and tasks. This correspondence approach, however, is bolstered by its explication of elements that can be readily specified, for instance, the types of rubrics and exemplar papers that can be used to illustrate a performance level. Most examples in educational reform at best approach the correspondence rather than the congruency level, although few are as thorough as they might be. As to the "harmony" implied in the definition, we are on weaker ground. Harmony within a discipline at a single grade, or up and down the grades, is rarely achieved. Harmony among the disciplines is rarely considered. Only in theory might we imagine the degree to which disciplines and areas of inquiry complement, support, or contrast with one another, or the extent of transfer of learning to be attained by

the study of structured or discursive knowledge. Yet, finding the explicit points of relationship in the form of corresponding illustration may be an attainable strategy for educational alignment.

What are the consequences of this approach?

Risks of Correspondence

- Distance between "correspondences" is too far or not intuitive
- Few examples will be widely accepted
- Generalization by teachers and students will be difficult
- Nonaligned components will be obvious
- Relationships are exaggerated
- Trajectories are difficult to map out

Mitigations

- A partial solution based on examples
- Relatively easy to develop additional examples, differing by only a few variables

Alignment as a Bridge

Correspondence specifies how x relates to y. A bridge as a metaphor for alignment suggests connection of another type. A bridge is a pathway that provides the connector from one location to another. What is the bridge from standards to assessment tasks? It is typically agreed to be the performance standards (or operational objectives describing accomplishments of students at various levels of proficiency). What is the bridge between measures and instruction? In traditional practice, it is the general specifications used to design the test, describing a general content and skill matrix. With assessments of constructed performance, it is the scoring rubric. The scoring rubric, with its description of needed prior knowledge, criteria for expression and analysis, and levels of expertise demonstrated for the award of particular points, presents an operational set of guidelines for the designer of instruction or the teachers themselves. By no means specifying chronology, the rubric lays out an explicit set of requirements to be met. The risk of this approach to instructional design is that the rubric design is arbitrary and does not represent useful or general conceptions of performance in the domain of interest. This is much the same complaint that was accorded early efforts to form a bridge from objectives

to instruction using task analysis. Gagné (1965) and his followers used a question as a device (what does the learner need to be able to do in order to perform . . .) to understand the prerequisites of tasks. The risk of this approach was that the analysis was dependent upon idiosyncratic views of the tasks or goals by the analyst, or that he or she held close a particularly comfortable instructional strategy. In other words, what a learner needs to be able to do depends upon your preferred sequence of teaching. Although the myriad task analyses generated in the last 40 years and the cognitive task analyses in the last two decades have rarely been subjected to empirical verification, they nonetheless have the benefit of being explicit and understood by many, and have the potential to be generalized to other similar tasks. In the ideal world, rubrics would be subjected to rigorous validity studies (see Baker, Freeman, & Clayton, 1991) or minimally to criteria promulgated to support high-quality measurement (see Linn, Baker, & Dunbar, 1991). Their great benefit as a bridge is that empirical data can be obtained to support their use.

Other key bridges need to be built, or at least reinforced. For example, one major weakness in the entire RBR edifice is the lack of good measures of classroom practice. If learning is the key condition, then how do we reconcile a system that uses second- or third-order indicators as means to understand what is occurring in classrooms? One approach developed at CRESST and empirically validated in LAUSD (Matsumura, Garnier, Pascal, & Valdés, 2002) is the analysis of assignments teachers give to students as a measure of the teachers' understanding simultaneously of the standards they are expected to teach, the levels of performance their students are expected to achieve, and their understanding of external measures. Looking at teacher assignments (validating them with student performance and work) is an excellent proxy for measuring instructional process, but our collective efforts in that realm are not yet convincing.

Obviously the strongest bridge possible, and one we hope to see more of, is a coherent curriculum, that is, one that exemplifies both broadly and concretely the intentions of the standards, and the content and skills to be taught and learned. Depending upon its development, the curriculum itself may be closer to a correspondence approach than a congruence model. But most important, it presents, even in syllabus form, an analytical and chronological support for instructional practice, classroom assessment design, and external measure development.

Risks of Bridges

- Weak in description and logical and empirical relationships
- Inadequate professional development to encourage use
- Weak prioritization

Mitigations

- Documented guidance against which progress can be measured

Alignment as Gravitational Pull

The Google search generated many examples of alignment related to scientific processes, most frequently astronomical phenomena. One metaphor provoked by this array of cosmic geometry is the idea of alignment as gravitational pull. Gravity, its force a product of mass moderated by distance, enables us to make precise predictions about the location of objects in time and space. Alignment of educational outcomes, processes, and goals could well be specified by looking at the centripetal forces that hold disparate activities together. For example, if standards, instruction, and assessments are designed to focus on a set of cross-curricular skills, such as problem solving, knowledge acquisition and understanding, communication, and metacognition (Baker, 2003b), then these common underpinnings, made explicit by definitions and examples, could be used to hold together otherwise disparate pieces of content. The elements, at their best, should be general across more than one (but certainly not every) subject matter domain. These elements, for instance, of problem solving, then can be reused again and again, permitting both the efficient and the effective design of learning. Each of the sets of cognitive demands, acting as the core features of reform, would have characteristics that delimit optimal teaching strategies, learning experiences, and measurement design and scoring. How change would occur depends on answers to the following questions:

- Into what content or subject matters are the domain-independent requirements embedded? For example, what does problem identification (a key element in problem solving) look like in algebra word problems, in prose analysis of novellas, or in determining unknown substances in chemistry?

- What features of the domain-independent problem-solving model (or that of knowledge understanding, etc.) are appropriate to create a developmental path of greater expertise? How do these features vary explicitly by grade level? (For example, how complex is the masking or

conflicting information provided to obscure the identification of the problem?)

- What elements of scientifically-based content models (of learning or of pedagogy) should be included in the subject matter goals, instruction, and outcome measurement, that is, in domain-specific models (Pellegrino, Chudowsky, & Glaser, 200l)?

The alignment, in this case, would be based on a common set of elements at the center of the reform, rather than the analysis of superficial features.

A second important source of gravitational pull, or the holding together of a system, is something outside the typical discussion of achievement-based educational reform. The development of social capital at the institutional (school) or organizational (district) levels can provide another axis on which the system can rotate. Social capital (see Hargreaves, 2003) involves a recognition of the shared priorities of a group, their beliefs in collective efficacy, trust, networking, and transparency (Baker, 2003a). Even in an environment with partial academic alignment, the alignment of motives and efforts will exert a powerful force on the local educational system. Such social alignment among individuals in an institution will give rise to the pursuit of the details that may be missed in a more bureaucratic approach to educational alignment.

The second notable attribute of the gravitational pull model is that relationships are dynamic. Systems move in and out of alignment based on instructional emphases, staff capability, and policy decisions. This approach should smooth out the differences and permit the system to be less volatile.

Risks of Gravitational Pull

- Requires sophisticated understanding of cognition and subject matter
- Changing priorities in subject matter may be viewed as a weakness

Mitigations

- Deep relationships can persist, supporting teaching and learning
- Transfer can be incorporated as a goal of learning
- Recognizes the dynamic and changing relationships in educational reform rather than imagining a steady state

I believe that the practice in educational reform does not need to exhibit a lockstep progression through these different metaphors, albeit that for now, we are locked into the congruency concept with occasional bows to correspondence

processes. Rather, the notion of common cognitive demands would allow the inevitable changes to be made in the topical content of the instruction without disturbing the fundamental relationships of the reform elements. This approach also presages the measurement and, if luck holds, the teaching of transfer, for accomplishment of the cognitive tasks can be implemented in examples that might share the same content standard (e.g., understanding the role of the frontier in American history), but vary greatly in context, format, and specific topic.

Measuring Alignment

An excellent article analyzing alignment by Bhola, Impara, and Buckendahl (2003) describes the processes to date used to characterize, document, and quantify alignment. The authors discuss levels of complexity of alignment, and focus for the most part on the alignment of standards to measures (or the reverse). They describe the important work of Webb, Porter, and others (e.g., Porter, 2002; Webb, 1997, 1999). They reference research by Herman, Webb, and Zuniga (2002), who used linguistic criteria as well as content, cognition, and task requirements in their study of alignment. Almost all of these studies flow from a congruency model, although Porter's work, emphasizing the theoretical sampling of a standard's domain, may have something of the correspondence metaphor in it.

R&D Priorities

Where do we go from here in the measurement, or at least the solid documentation, of alignment? Clearly the area that needs most attention is the measurement of instructional and learning processes. We need to know what is actually happening in classrooms, what work students are being given, and how teaching and learning are taking place. The current approaches, involving questionnaires answered by students and teachers, observations by experts or peers, and logs or other chronological records, suffer from many limitations. Accuracy, reliability, validity, feasibility, and cost are just a few, but total to the claim that we have no scalable approaches for understanding what is happening in instructional practice. A black box for the key active component of schooling is unacceptable, especially if direct practice of test items or item-like events is to be avoided. Approaches to documenting alignment in a regular, scalable, and accurate way will undoubtedly require a switch in our general strategy of declaring alignment as a proportion of all or nothing.

Rather, close-up studies must be continued that address the extent to which different instructional strategies have differential effects on measures, or to shift the focus, the extent to which different measures have known propensities to respond to instructional treatments (see Niemi, Baker, Steinberg, & Chen, 2004).

Once there is evidence of instructional approaches that work, there can be research undertaken to target the measurement of those processes in classroom settings. Such a state of affairs would be far more satisfying than a census listing of what had been "covered" in content. Moving from measures of coverage to engagement to learning should be our goal.

Among the vast number of unresolved problems, which might we emphasize and select for first priorities?

Earlier I made reference to the LEARNOME as a research approach, and I will expand only briefly. Whatever its components—linguistic requirements, explicit maps of prior and to-be-learned content knowledge, families of cognitive demands, task sequences—investigations into such an area will force the development of missing parts of our reform goals. First, what language (lexicon) can we use to identify the pieces that we casually describe using our own implicit definitions? Can we develop a common language to describe educational processes and outcomes? Second, at what level of granularity should we engage—how micro should the relationships be between learning and measurement? If the research is to be trusted at all, the relationships need to be made at very detailed levels. Third, how can we predict performance based on these relationships? For whom and with what accuracy? How can these relationships, at a system or curriculum level, or even as the learning of a particular child, be represented? How can a test or instructional segment be characterized so that its shared and disparate elements are obvious? Can technology provide us with an automated analysis of the relationships among performance standards, tests, classroom assessments, texts, and student activities?

Can we conduct validity studies of "power" goals or standards, those outcomes that either provide underpinnings for a wide range of outcomes or require a known set of steps to be achieved prior to accomplishment? If we can identify these power standards, we can vastly simplify the process of alignment and move to a more focused set of standards as recommended by Popham and his colleagues on the Commission on Instructionally Supportive Assessment (2001).

Given the turbulence in policy, a desirable goal would be to find a way to simulate what would happen if student learning were stable over time, sheared of annual policy accoutrements. As of now, we have only a vague idea of what continuity of learning content and method would feel like from the child's eye up.

On the development side of R&D, we can support alignment by creating assessment design tools for use by classroom teachers and by external assessment developers. If based on common constructs and operational definitions, teachers' classroom analyses could make a deeper contribution to the view of educational effectiveness we have developed.

Summary

Now to state the obvious. Useful metaphors are rarely made operational. Alignment should be treated as a goal rather than as a bureaucratic requirement to be checked off. If it is the latter, we will never attain it. And should we believe we have "aligned" our system, we must remember that the world moves, and alignment strengthens and weakens with change. In addition, we need more powerful conceptual analyses and tools to achieve reform ends. Let us find the centripetal elements that hold systems together—cognitive demands that can be taught, learned, measured, and transferred—and the social capital that motivates and energizes effort—or identifiable shared experience. Only a system that is held together from the inside can stimulate the deep and useful learning of all children.

References

America 2000 Excellence in Education Act. (1991). Proposed legislation (H.R. 2460), 102d Cong. 1st Sess. (ERIC Document Reproduction Service No. ED341115)

Baker, E. L. (1972). *Preparing instructional materials* (Report to the U.S. Department of Health, Education, and Welfare, Office of Education, Project Number 1-0027, Grant Number OEG-0-71-0645). Los Angeles: University of California, Graduate School of Education. (ERIC Document Reproduction Service No. ED 068458)

Baker, E. L. (1973). The technology of instructional development. In R. M. W. Travers (Ed.), *Second handbook of research on teaching* (pp. 245-285). Chicago: Rand McNally.

Baker, E. L. (2003a). *From usable to useful assessment knowledge: A design problem* (CSE Rep. No. 612). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Baker, E. L. (2003b). Multiple measures: Toward tiered systems. *Educational Measurement: Issues and Practice*, *22*(2), 13-17.

Baker, E. L., Freeman, M., & Clayton, S. (1991). Cognitive assessment of history for large-scale testing. In M. C. Wittrock & E. L. Baker (Eds.), *Testing and cognition* (pp. 131-153). Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., & Linn, R. L. (2004). Validity issues for accountability systems. In S. Fuhrman & R. Elmore (Eds.), *Redesigning accountability systems for education* (pp. 47-72). New York: Teachers College Press.

Baker, E. L., & Popham, W. J. (1973). *Expanding dimensions of instructional objectives.* Englewood Cliffs, NJ: Prentice-Hall.

Baker, E. L., Sawaki, Y., & Stoker, G. (2002, April). *The LEARNOME: A descriptive model to map and predict functional relationships among tests and educational system components.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Bhola, D. S., Impara, J. C., & Buckendahl, C. W. (2003). Aligning tests with states' content standards: Methods and issues. *Educational Measurement: Issues and Practice*, *22*(3), 21-29.

Commission on Instructionally Supportive Assessment (W. J. Popham, Chair). (2001, October). *Building tests to support instruction and accountability: A guide for policymakers.* Convened by: American Association of School Administrators, National Association of Elementary School Principals, National Association of Secondary School Principals, National Education Association, National Middle School Association. Retrieved April 2, 2004 from http://www.aasa.org/issues_and_insights/assessment/Building_Tests.pdf.

EFF Assessment Consortium: SRI International and Center for Literacy Studies. (2004, January). *Equipped for the Future. EFF standards and performance level descriptors for: Reading, writing, math, speaking, and listening* (Prepared for the National Institute for Literacy). Knoxville: University of Tennessee, Center for Literacy Studies. [See also National Institute for Literacy, Washington, DC, http://www.nifl.gov/lincs/collections/eff/eff.html]

Elementary and Secondary Education Act of 1965, Pub. L. No. 89-10, 79 Stat. 27 (1965).

Gagné, R. M. (1965). *The conditions of learning*. New York: Rinehart and Winston.

Goals 2000: Educate America Act, Pub. L. No. 103-227, 108 Stat. 125 (1994).

Hargreaves, D. (2003, January). *From improvement to transformation.* Keynote address to the International Congress for School Effectiveness and Improvement 2003 conference "Schooling the Knowledge Society," Sydney, Australia.

Hauser, R. (2004, March). High stakes. In *Old lessons for a new decade: Reflections on BOTA projects.* Presentation at the meeting of the Board on Testing and Assessment, National Research Council, Washington, DC.

Herman, J. L., Webb, N., & Zuniga, S. (2002). *Alignment and college admissions: The match of expectations, assessments, and educator perspectives.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Holland, J. G., & Skinner, B. F. (1961). *The analysis of behavior: A program for self-instruction.* New York. McGraw-Hill.

Improving America's Schools Act of 1994, Pub. L. No. 103-382, 108 Stat. 3518 (1994).

Kelly, H. (2004, February). *Integrating technology with cognitive science to improve assessment and learning*. Presentation at the American Association for the Advancement of Science Annual Meeting, Seattle, WA.

Koretz, D. (2003a). *Teachers' responses to high-stakes testing and the validity of gains: A pilot study* (Draft deliverable to the National Center for Research on Evaluation, Standards, and Student Testing). Cambridge, MA: Harvard Graduate School of Education.

Koretz, D. (2003b, April). Using multiple measures to address perverse incentives and score inflation. In *Multiple perspectives on multiple measures.* Symposium presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15-21. (ERIC Document Reproduction Service No. EJ 436999)

Lumsdaine, A. A., & Glaser, R. (Eds.). (1960). *Teaching machines and programmed learning: A source book*. Washington, DC: National Education Association of the United States.

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, Part II* (pp. 104-140). Chicago: University of Chicago Press.

Matsumura, L. C., Garnier, H., Pascal, J., & Valdés, R. (2002, April). *Classroom assignments as indicators of instructional quality*. Presentation at the annual meeting of the American Educational Research Association, New Orleans, LA.

Morris, W. (1981). *American heritage dictionary of the English language*. Boston: Houghton Mifflin.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform*. A report to the nation and the Secretary of Education. Washington, DC: U.S. Government Printing Office.

National Council on Education Standards and Testing. (1992). *Raising standards for American education. A report to Congress, the Secretary of Education, the National Education Goals Panel, and the American people*. Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1991a). *Measuring progress toward the national education goals: Potential indicators and measurement strategies* (Compendium of interim resource group reports). Washington, DC: Author.

National Education Goals Panel. (1991b). *The National Education Goals report, 1991: Building a nation of learners*. Washington, DC: Author.

National Education Goals Panel. (1991c). *Potential strategies for long-term indicator development: Reports of the technical planning subgroups*. Washington, DC: Author.

National Education Goals Panel. (1992). *The National Education Goals report, 1992: Building a nation of leaders*. Washington, DC: U.S. Government Printing Office.

National Education Goals Panel. (1993). *Report of goals 3 and 4, Technical Planning Group on the Review of Education Standards*. Washington, DC: Author.

National Governors Association. (1991). *Results in education, 1989: The governors' 1991 report on education*. Washington, DC: Author. (ERIC Document Reproduction Service No. ED 313338)

Niemi, D., Baker, E. L., Steinberg, D. H., & Chen, E. (2004, April). Validating a large-scale performance assessment development effort. In J. Evans (Chair), *Applying research-based performance assessment models in routine practice in a large urban school district: The pleasure-pain principle*. Symposium presented at the annual meeting of the American Educational Research Association, San Diego.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, 115 Stat. 1425 (2002).

Pellegrino, J., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessments* (Committee on the Foundations of Assessment; Board on Testing and Assessment, Center for Education. Division on Behavioral and Social Sciences and Education). Washington, DC: National Academy Press.

Popham, W. J., & Baker, E. L. (1970). *Systematic instruction*. Englewood Cliffs, NJ: Prentice-Hall.

Porter, A. C. (2002). Measuring the content of instruction: Uses in research and practice. *Educational Researcher, 31*(7), 3-14.

Sawaki, Y., Xi, X., Stoker, G., & Lord, C. (2004). *The effects of linguistic features of NAEP Math items on test scores and test-taking processes* (Draft deliverable to U.S. Department of Education). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Tyler, R. W. (1949). *Basic principles of curriculum and instruction*. Chicago: University of Chicago Press.

Webb, N. L. (1997). *Research Monograph No. 6: Criteria for alignment of expectations and assessments in mathematics and science education.* Madison, WI: National Institute for Science Education.

Webb, N. L. (1999). *Research Monograph No. 18: Alignment of science and mathematics standards and assessments in four states.* Madison, WI: National Institute for Science Education.