DRAFT October, 2000

Are Content Standards Being Implemented in the Classroom? A Methodology and Some Tentative Answers

Andrew C. Porter and John L. Smithson

Arguably the most notable trend in education policy in the past ten years has been the movement toward a standards-based approach to insuring the quality of education provided to all children. Standards have been set by professional organizations, such as the National Council of Teachers of Mathematics (NCTM), the American Association for the Advancement of Science (AAAS), and the National Council of Teachers of English (NCTE), by the states and strongly encouraged by the federal government. Title I of the ESEA legislation requires all states to adopt challenging content and performance standards in at least reading, language arts and mathematics.

The question is whether standards-based reform is making a difference in the type and/or quality of instruction experienced by students. This chapter focuses on the issues that must be addressed and the challenges that must be overcome to provide a credible answer to questions regarding the impact of standards on the quality of instruction received by students. Results from some preliminary investigations are reported.

Determining the impact of standards on classroom practice can be viewed as a three-part problem. First, one needs a description of the relevant educational practice that permits comparison to the standard or goal being targeted. Second, one must establish the target (i.e. just what *are* the standards of concern, and how will one know if they have been met?). Third, it is necessary to have an explanatory model by which to attribute the practice described as a result of

---

Andrew C. Porter is the Director of the Wisconsin Center for Education Research, Director of the National Institute for Science Education, and a Professor of Educational Psychology at the University of Wisconsin-Madison.

the standards established.

Establishing causal relations between indicators of education processes and school outputs is complicated, and the results always tentative, especially from correlational studies such as an indicator system would support. Although there are differences of opinion about how useful such analyses can be in diagnosing the relative utility of different types of educational practices, most agree that such indicator data are better than no information at all.

Because the number of potential school process variables is large, some criteria are needed for deciding which to measure. If what is wanted is an index of opportunity to learn, then the criterion for establishing priority should be utility for predicting gains in student learning (achievement). The best predictors of student achievement gains are the properties of instruction as it occurs in schools, what content is taught, how effectively, to which students, and to what levels of achievement. Our discussion will focus on descriptions of instructional content, as these descriptions seem best suited to explaining student achievement.

In what follows, a framework for attributing instructional practices to standards-based reform is offered as a context for examining how such analyses might proceed. Examples are given of recent work describing the content of practice, followed by a discussion on determining the content implications of standards-based policy instruments. Procedures for measuring alignment are developed and illustrations given for assessing alignment between policy instruments (e.g. standards, assessments) and instruction, as well as illustrations of alignment between instruction and gains in student achievement. The chapter concludes with consideration of the issues that must be addressed in measuring the content of instruction.

John Smithson is an Assistant Researcher and Project Director at the Wisconsin Center for Education Research.

Attributing Causality

While it is true that education presents an exceptionally complex system with numerous steps in the causal chain between policy tool and student effects (Kennedy 1999), for the purpose of this discussion we simplify the causal chain into three key components: the *intended* curriculum, the *enacted* curriculum, and the *learned* curriculum (i.e., student outcomes). The logic behind this chain of causality suggests that the *intended* curriculum --as represented by policy tools such as content standards, curriculum frameworks/guidelines and state assessments-- influences teacher practice (the *enacted* curriculum), which in turn impacts student learning as measured by state assessments. The necessary evidence to attribute changes in student outcomes to policy initiatives can be divided into two parts. One part of such an explanation is to provide evidence that policies have changed practice in desirable ways, while the second part of the explanation seeks to make the link between practice and outcomes. Both parts are necessary in order to draw the link between policy initiative and student achievement. Kennedy's critique on this point is that researchers tend to focus upon one or the other. That researchers focus on one or the other explanation is perhaps not surprising, as each of the two explanatory pieces to the overall causal puzzle require different types of evidence, reasoning and theory.

Linking the intended and enacted curricula

Assuming for the moment that one has comparable, quantifiable descriptions of the intended curriculum or 'target' and the enacted curriculum, a measure of agreement or alignment between the intended and enacted curricula can be calculated. Such an alignment measure is

much like a correlation in that it suggests a relationship but is insufficient to support a causal connection. Thus, if the resulting alignment measure was high (indicating strong agreement between the intended and enacted curricula), one would still need more information, as well as a theoretical framework, to argue causality. For example, one useful additional piece of information would be longitudinal data indicating change in practice over time. If such data were available, and those data indicated the direction of change over time was towards greater alignment with the intended curriculum, one would have a stronger case, but still insufficient to be confident about cause and effect. In addition, one needs a theoretical model in which to set the explanation for policy influence.

An example of one such model has been offered by Porter (1991). In this model, policy tools are described on the basis of four characteristics: prescriptiveness, consistency, power and authority. Prescriptiveness indicates the extent to which a policy instrument specifies desired practice. Consistency describes the extent to which policy instruments are mutually reinforcing (i.e. aligned). One important measure of consistency is the extent to which the content standards and assessments of a given state present a common message about the intended content of instruction. A curriculum policy instrument has power to the extent that rewards and sanctions are tied to compliance with the policy. High stakes tests are one notable example of a curricular policy with power. Authority refers to the extent to which policies are persuasive in convincing teachers that the policy is consistent with notions of good practice.

The hypothesis is that the more a curriculum policy reflects these four characteristics the stronger the influence that policy will have on curricular practice. Thus if a specific policy or set of policies is shown to be strong on several or all of these characteristics, if descriptive data

reveal substantial agreement between descriptions of the intended and enacted curricula, and if this level of agreement has increased over time as the policy has had an opportunity to exert an influence, one can begin to make claims of attribution.

Linking the enacted curriculum to student outcomes

Having sufficient evidence to attribute instructional practice to the influence of policy instruments (such as content standards and state assessments) still falls short of explaining student outcomes. To stretch the causal chain to include outcomes, evidence is necessary to make the link between instructional practice and gains in student learning. Note here the reference to 'gains' rather than achievement. While achievement scores alone provide some indication of the level of knowledge students have attained, they say nothing about when and how that knowledge was acquired. To measure the contribution of instructional practice to a student's score a more narrow measure of achievement is necessary. By focusing on gains in student achievement, rather than simple test scores, it is possible to examine the contribution of classroom experience to student achievement over specified periods of time. This is essential if one is trying to demonstrate the effects of recent changes in policy and instruction on achievement (Meyer, 1997).

In addition to controlling for prior achievement (accomplished by the use of learning gain measures), one must also control for the socioeconomic status (SES) of students' families. In a recent analysis of results from the *Prospects* study (a large-scale nationally representative study), Brian Rowan (1999) found that prior achievement and SES accounted for as much as 80% of the variance in mean achievement among classrooms. Rowan estimated the percentage of variance among classrooms to be 11% after controlling for prior achievement and SES. This suggests that

the extent to which the classroom experience of students in a given year contributes to their overall achievement score is relatively small compared to these other factors. However, Rowan also notes that the percentage of variance attributable to classroom differences may be significantly higher when the alignment between the test and instruction is taken into account.

If one has comparable descriptions of the content of instruction and the assessment being utilized, an alignment variable can be calculated. This measure, used in conjunction with controls for prior achievement and SES, may be suitable for attributing achievement gains to instruction, particularly if alignment succeeds in predicting student achievement above and beyond the control variables.

Recent Efforts at Describing Instructional Practice

The largest, and best-known effort of recent years to describe instruction has been the Third International Mathematics and Science Study (TIMSS). Though perhaps best known for its national rankings of student achievement scores on the TIMSS assessments, substantial information was also collected on instructional practice, using teacher surveys and video-taped observations of classroom practice (Peak 1996). Analyses of textbooks conducted as part of the TIMSS revealed a characteristic of mathematics and science instruction in the United States that has since become a familiar refrain among math and science educators. The description of mathematics and science curricula in the U.S. as being "a mile wide and an inch deep" is now commonly offered as an explanation for the mediocre performance of U.S. students on the TIMSS assessments (Schmidt, et.al. 1999).

Another example of the uses to which descriptions of classroom practice have been put is

provided by the *National Evaluation of the Eisenhower Professional Development Program*, in which researchers conducted a longitudinal study of the effects of professional development activities on teacher practice in the classroom. In this study researchers used descriptions of classroom practice and instructional content collected over a period of three years to track changes in practice as a result of professional development. Results from this study suggest that professional development activities with a clear content-focus lead toward increased emphasis on those topics during instruction. Researchers identified several other characteristics of professional development that also appear effective in changing teacher practice. These included the use of active learning strategies as part of the professional development activity, collective participation by a group of teachers from the same school or grade level, linking professional development opportunities to other activities, and designing activities that build upon teachers' prior knowledge (Porter, et.al. in review).

A third example concerns the development of a set of survey instruments as part of a multi-state collaborative to provide a set of practical tools for collecting consistent data on mathematics and science teaching practices and instructional content. The surveys define a comprehensive set of indicator data on instructional processes for elementary, middle and high school classes in mathematics and science. Both teacher and student surveys have been developed. The teacher surveys for each subject and grade level consist of two distinct instruments. One instrument (Survey of Instructional Practices) focuses upon instructional activities, teacher opinions, characteristics and background, professional development, school and student characteristics. The other instrument (Survey of Instructional Content) collects detailed information on content using a two-dimensional content matrix design developed by

Porter and Smithson and based on previous work (Porter, et.al 1988; Porter, et.al. 1993). Together these instruments are referred to as the Surveys of the Enacted Curriculum (SEC). The SEC instruments are currently being employed in at least three separate studies of mathematics and science reform. One of these studies (funded by the eleven participating states and the National Science Foundation) is exploring the efficacy of the SEC instruments as a tool for states and researchers to use in monitoring reform and in evaluating the efficacy of reform efforts on classroom practice (CCSSO 2000). Data collected from this eleven-state study indicate differences in teaching styles and instructional content that emerge between grade levels and between teachers that are and are not involved in pursuing reform strategies. The results are being shared with states and the participating schools, in order to provide them descriptive information about practice, professional development, and teacher opinions for use in monitoring reform and evaluating improvement efforts.

In addition to this state initiated study, two independent evaluations of the urban systemic initiatives program (USI) are being conducted utilizing the SEC instrumentation. Both studies, one conducted by Systemic Research Inc., *How Reform Works: An Evaluative Study of NSF's Urban Systemic Initiatives* (Ware, et.al. 2000), and the other by a team of researchers at the University of Southern Florida, *Assessing the Impact of the National Science Foundation's Urban Systemic Initiative*, are using the SEC instruments in conjunction with in-depth case studies, observations and interviews to examine the effectiveness of USI programs currently funded in forty urban school districts (Blank, et.al. 2000). The use of a standard set of instruments across research studies, such as is now occurring in these three studies, provides a unique opportunity to examine descriptions of practice across studies.

Selecting the Target

Though an essential piece of the educational puzzle, descriptions of practice alone are insufficient to determine the extent to which standards are being implemented. What is necessary is some target against which to compare those descriptions. This 'target' needs to be set in terms of what one should see happening in classrooms. Ideally, the target would be described using the same language used to describe practice. There are at least three potential targets:

One obvious choice for a target would be a state's content standards, or possibly the content and pedagogy standards set forth by one or another professional association (e.g., NCTM). The challenge with such sources is translating the language contained in documents describing the standards into a clear picture of desired classroom practice. Many states have vague and visionary statements of practice incorporated into the language of their standards. This leaves a good deal of room for interpreting just how instruction should look in a specific classroom on a day to day basis. If the target is fuzzy, determining the extent of standards implementation will also be fuzzy.

For this reason, one might turn instead to curriculum guides or frameworks. These tend to be more specific than standards, though curriculum guides also leave considerable ambiguity as to exactly what unit is to be taught when. As with standards, the more prescriptive the curricular materials are the better (for the purpose of content analysis and according to our theory).

Yet another potential source for establishing a description of the 'target' against which to

compare practice are state assessments. Assessments have the advantage of presenting clear indications of what content is considered important, as well as the level of knowledge expected of students with regard to that content. In that sense assessments are prescriptive, in that they specify particular topics and the depth of knowledge considered most important. Perhaps more importantly, since student outcome measures are often the basis for rewards and punishment (whether for students, teachers or schools), using the assessment as a target can be useful in diagnosing why students succeed or fail on assessments.

On the other hand, assessments are not prescriptive in the sense of defining well what *should* be taught, since the items on an assessment represent only a sample from the content domain the assessment is intended to represent. As a result, one problem with using assessments to describe the intended curriculum is that any particular form of an assessment will necessarily represent only a sample of items from the domain of interest. Thus, where multiple test forms exist, content analyses of the assessments should include all items across all forms in order to capture a more complete picture of the content message embedded in the assessment instruments.

Ideally, assessments will be 'aligned' to the standards; i.e., they are intended to convey the same content message as implied by the standards. With descriptions of practice, and comparable descriptions of the 'target' for successful standards implementation in hand, it is possible to measure the degree of alignment between instruction and the target. The higher the degree of alignment between instruction and the target, the greater the extent to which standards can be said to have been implemented in the classroom.

Of course measured for only a single point in time, alignment is only correlational, and thus not strong evidence of cause and effect (though low levels of alignment make clear that

standards are not being fully implemented). Stronger evidence can be developed through investigation into whether alignment increases over time after standards have been put in place.

Measuring Alignment between Assessments and Instruction

As part of the eleven-state study mentioned earlier, six states participated in a sub-study to analyze alignment between instruction and assessments. To the extent that instruction is aligned to a state's assessment, one link in the complex causal chain necessary to connect policy initiatives and student outcomes can be established. Presumably, if standards-based reform is having an effect, instruction in a state will be more aligned to that state's test than to tests given by other states.

The assessments analyzed were mathematics and science tests in grades 3,4, or 5 at the elementary level, and grades 7 or 8 at the middle school level (depending upon the grade level assessed in a given state). The majority of assessments analyzed were for grades 4 and 8, which coincided with the grade level at which teacher descriptions of practice were collected. For some states, multiple forms were analyzed (which was the goal). In addition, NAEP math and science assessments were content-analyzed. All grade 4 and 8 NAEP test items were included in the content analysis of the NAEP assessments.

Tests were content-analyzed, item by item, using the same language and distinctions for describing content (topics by cognitive demand) as employed in the teacher surveys described earlier. Six state mathematics representatives, six state science representatives, three university math educators and four science educators were involved in the content analyses that were conducted during a two-day period in the summer of '99.

Descriptions of practice were based upon survey results from 503 teachers across eleven states. These included elementary and middle school teachers reporting on their science or mathematics instruction for the then current school year (the surveys were administered in the spring of 1999) using the SEC instruments. It should be noted that the participating teachers do not offer a representative sample, particularly at the state level, as sampling was neither random nor sufficiently robust to warrant generalizations. Therefore the alignment measures described below are presented for illustrative purposes only.

These data allow investigation of assessment-to-assessment alignment (including state assessment alignment with NAEP), instruction-to-assessment alignment, and instruction-to-instruction alignment (state by state), at each grade level for each subject. For each test, the average degree of emphasis on a topic (e.g. linear equations) by cognitive demand (e.g. solve novel problems; see Table 6) intersection, across content analyzers, was calculated. The result was a matrix of proportions, with dimensions topic by cognitive demand. Similar topic-by-cognitive demand matrices of content emphasis were calculated for instructional content based on teacher reports. An alignment index with a range from 0 to 1 was created and calculated to describe the degree of alignment between assessments and instruction.* As can be seen from Table 1, state tests were generally more aligned with each other (.32 to .45) than they were with NAEP (.24 to .34), though the differences are not large. For each subject and grade level, state test to state test alignment is, on average, higher than is state test to NAEP alignment. At the same time these data establish that each state test presents a unique target for instruction. State

* Alignment between pairs of topic by cognitive demand matrices of content emphasis were calculated as, $I = 1 - (\Sigma|x-y|)/2$, where x is the cell proportion for one matrix (representing an assessment or instruction) and y is the corresponding cell proportion for the other matrix (representing an assessment or instruction).

tests are not interchangeable.

Instruction in a state was, in general, no more aligned to that state's test than it was aligned to the tests of other states, suggesting that standards-based reform has not yet brought instruction into alignment with these states' tests. In mathematics, instruction was more aligned with NAEP than with state tests. In science, the opposite was true (See Table 2). Two caveats are important. First, recall that the data on instruction is illustrative only. The samples of instruction from each state cannot be taken as representative of that state, as the samples are neither random nor sufficient in size for such inferences. Secondly, to the extent that a state's test is not aligned to the state's content standards, one might not want instruction to be highly aligned to the state test. For example, if the state assessment is a basic skills tests, one would hope that instruction would still get beyond the basics. Nonetheless, to the extent a state test is used in an accountability program, it may have an influence on instructional practice. While only illustrative, these analyses and results do provide some indication of the utility that such measures would hold if based on a more representative sample.

The data suggest that instruction in one state is quite similar to instruction in another state. Average instruction-to-instruction alignment indicators ranged from .63 to .80 (see Table 3). However, one should be careful not to interpret this as an indication that there is little variation in practice across teachers. When individual teacher reports of content are compared within a state, and even within a school, the degree of alignment drops considerably.

It is also possible to use the alignment index to measure the level of inter-rater agreement in the content analysis of the test items. Illustrative results are reported for elementary and middle school math and science (see Table 4). In interpreting these inter-rater agreements, it is

important to realize that any one item may assess several different types of content. Raters were limited to selecting only three topic-by-cognitive demand combinations per item. This undoubtedly forced some disagreements among raters. When making distinctions at the finest grain (i.e. topics by cognitive demand), alignment is in the neighborhood of .40 to .50. Alignment is obviously better when comparisons are made at a larger grain size (such as algebra or geometry) by cognitive demand. Since assessments were described as the average across raters, and each form was content analyzed by at least four experts, the validity of the descriptions of the tests is high.

Recent Efforts at Linking Instruction to Student Outcomes

What about the second link in the causal chain? At least two recent studies, one in mathematics (Gamoran, et.al. 1997) and one in science (Wang 1998) have utilized sophisticated quantitative modeling tools (Bryk and Raudenbush 1992) to demonstrate the power of classroom measures of instructional content in predicting achievement gains among students. Because the mathematics study utilized an approach to instructional alignment similar to the one discussed above, it is described in more detail here.

The study looked at the efficacy of transition math courses in California (Math A) and New York (Stretch Regents) in bridging the gap between dead-end, basic math courses and college preparatory courses for low achieving, low income students. The sample consisted of two schools from each of four districts, two districts in California, and two districts in New York. Two schools were then selected from each of the four districts. Within each school, the sample included at least one traditional low-level course (e.g., general math or pre-algebra) and at least

one college preparatory course (e.g., Regents One, algebra, or geometry). In one district, all lower-level mathematics courses had been eliminated, representing an exception to this design. In total, 56 classrooms from seven schools participated in the study.

Using a combination of survey, observation, interview and pre/post test data, researchers constructed comparable descriptions of practice and assessment content in order to describe the relationship between different course-types. Since all students in the participating classes were administered the same test (drawn from public release NAEP items) it was also possible to look at gains in student achievement across the differing course-types to compare the differential effects of instruction in the college-prep, transition, and basic math courses.

Several indicators were formed for investigating the relationship between the content of instruction delivered in the classroom as reported by the teacher, and student achievement gains on the achievement test constructed from NAEP public release items. One indicator of content coverage of tested material is the proportion of instructional time spent covering tested content (level of coverage). Another indicator is the match of relative emphases of types of content between instruction and the test (configuration of coverage). The mean proportion of instructional time that addressed one or more types of content tested (level of coverage) was .07, with a standard deviation of .02. The mean for configuration was .58, with a standard deviation of .08.[†]

As can be seen in Table 5, alignment can be defined based on topics (the rows in the content matrix), or cognitive demand (the columns), or at the intersection of topics by cognitive

† To form the configuration of coverage index, first the proportion of instructional time for each type of content tested relative to the total amount of instructional time spent on tested content was determined. Second, the proportion of test items in each tested area was determined. The absolute value of the difference between these proportions was summed, and the sum re-scaled by dividing by 2.0 and subtracting the result from 1.0, to create an

demand (the cells in the matrix). The highest correlations with student achievement gains arise when content is defined at the intersection of topics by cognitive demand. The correlation for class gains is .48 and for student gains is .26. These are substantial correlations for predicting student achievement gains.

Issues in Defining Indicators of the Content of Instruction

There are several problems that must be solved in defining indicators of the content of instruction.

Do We Have the Right Language?

Getting the right grain size. One of the most challenging issues in describing the content of instruction is deciding upon the level of detail of description that is most useful. Either too much or too little detail presents problems. For example, if description were at the level of distinguishing math from science, social studies, or language arts, then certainly all math courses would look alike. Nothing would have been learned beyond what was already revealed in the course title. On the other hand, if content descriptions make distinctions which essentially identify the particular exercises on which students are working, then surely all mathematics instruction would be unique. At that level of detail, no two courses with the same title cover the same content because trivial differences are being distinguished.

An issue related to grain size is how to describe instruction that does not come in nice, neat, discrete, mutually exclusive pieces. A particular instructional activity may cover several topics and involve a number of cognitive abilities. The language for describing the content of

---

index which ranged from 0.0 to 1.0.

instruction must be capable of capturing the integrated nature of scientific and mathematical thinking.

Getting the right labels. The labels that are used to denote the various distinctions being made when describing the content of instruction are extremely important. Ideally, labels can be chosen that have immediate face validity for all respondents, so that questionnaire construction requires relatively little elaboration beyond the labels themselves. What is needed in order to have valid survey data is instrumentation where the language utilized has the same meaning across a broad array of respondents.

Some have reviewed our languages and suggested that the terms and distinctions should better reflect the reform rhetoric of the National Council of Teachers of Mathematics' mathematics standards (NCTM 2000) or the National Research Council's science standards (NRC 1996). But the purposes of the indicators described here are to characterize practice as it exists, and to compare that practice to various standards. For those purposes a reform-neutral language is appropriate. Still, one might argue that the language described here is not reform neutral but rather status quo. Ideally the language utilized should be translatable into reform language distinctions, so comparison to state and other standards is possible.

Another criterion for determining the appropriateness of the content language is to ask educators. As instruments have been piloted with teachers, the feedback has been surprisingly positive. Teachers often found completing the questionnaires to be an engaging, although challenging, task. They report that all too rarely are they engaged in conversations about their goals for instruction and the content of their enacted curriculum. When teachers have been provided descriptions of their practice (as in the case of the *Reform Up Close* study), these

descriptions have caused them to reflect. In many cases, they were surprised by what they saw, despite having provided the data themselves. This is probably a function of the analytic capacity of the language to describe not only what is taught and with what relative emphasis, but also what is not taught. As described earlier, these procedures are being offered by the Council of Chief State School Officers' State Collaborative on Assessment and Student Standards as tools for schools to use in self-study and reflection; tools which they believe will help schools pursue their reform agenda. Once again, educators are attracted to the kinds of data these types of instruments can provide.

Getting the right topics. Do we have content broken up into the right sets of topics? An alternative framework is in its beginning stages of development, under the auspices of the Organization for Economic Co-operation and Development (OECD) as part of their plan for a new international comparative study of student achievement. In that framework, big ideas are distinguished, (e.g., chance, change and growth, dependency and relationships, and shape). Clearly, this is a very different way of dividing up mathematical content than that taken here, and a very interesting one. Still, if the goal is to create a language for describing practice, practice is currently organized along the lines of algebra, geometry, measurement, etc., not in terms of big ideas. Perhaps practice should be reformed to better reflect the big ideas, but that has not happened yet.

Getting the right cognitive demand. When describing the content of instruction with a goal of building an indicator with a strong predictive value for gains in student achievement, content must be described not only by the particular topics covered (e.g., linear algebra, cell biology), but also by the cognitive activities which students are to be engaged in with those

topics (e.g., memorize fact, solve real-world problem). A great deal of discussion has gone into how many distinctions of cognitive demand should be made, what the distinctions should be, and how they should be defined. In the earliest work focusing on elementary school mathematics, just three distinctions were made: conceptual understanding, skills and applications (Porter, et.al. 1988). In the Reform Up Close study of high school mathematics and science (Porter, et.al. 1993), nine distinctions were made for both mathematics and science: (1) memorize facts/definitions/equations; (2) understand concepts; (3) collect data (e.g., observe, measure); (4) order, compare, estimate, approximate; (5) perform procedures: execute algorithms/routine procedures (including factoring, classify); (6) solve routine problems, replicate experiments/replicate proofs; (7) interpret data, recognize patterns; (8) recognize, formulate, and solve novel problems/design experiments; and (9) build and revise theory/develop proofs.

Later, in building yet another generation of the survey instruments (with funding from the National Center for Education Statistics), student cognitive activities were defined by the categories: (1) memorize: facts, definitions, formulas, (2) understand or explain concepts/ideas, (3) complete computations, follow detailed instructions, (4) solve equations that are given, (5) solve routine problems (e.g., stories/word problems), (6) solve novel/non-routine or real-world problems, (7) design experiments/empirical investigations, (8) collect, analyze, and/or report on data, (9) building/revise theory, develop proofs, and (10) explain solutions/answers to any type of problem.

Then, in the National Evaluation of Eisenhower, performance goals for students were defined as: (1) memorize; (2) understand concepts; (3) perform procedures; (4) generate questions/hypotheses; (5)collect, analyze, and interpret data; and (6) use information to make

connections. Each of these distinctions were further defined using descriptors. For example, "generate questions/hypotheses" had as descriptors: brainstorm, design experiments and solve novel/non-routine problems. "Use information to make connections" had elaborated under it: use and integrate concepts, apply to real-world situations, build/revise theory and make generalizations (see Table 6). The goal is to have distinctions on a questionnaire that are understood in the same way by each respondent. Obviously, with the types of distinctions made for cognitive demand, perfect clarity is not achievable.

One language or several? A related issue is whether a different language for describing the content of instruction is needed within a subject area at different grade levels or within a grade level for different subjects. In the *Reform-Up-Close* study (Porter, et.al. 1993), cognitive activities were described the same for both high school math and high school science. Obviously, the topics differed between mathematics and science and were largely unique for each subject area. When describing the content of elementary school instruction, however, many of the mathematical or science topics are different from the mathematical or science topics in high school courses. Different grade levels may require different languages for describing the content of instruction. It also may be that when describing the content of language arts instruction or social studies instruction, the cognitive activities and instructional mediums will be sufficiently distinct from those in math and science as to require new languages.

The possibility of a third dimension. Throughout the development of questionnaires to survey teachers on the content of their instruction, a third dimension to the content matrix has been entertained. In Reform Up Close, this third dimension was referred to as mode of presentation. The distinctions were: exposition-verbal and written; pictorial models; concrete

models (e.g., manipulatives); equations/formulas (e.g., symbolic); graphical; laboratory work; fieldwork. At various times, differing categories of modes of presentation have been tried. However, mode of presentation has not been a powerful addition to the descriptions provided by topics and cognitive demand. Mode of presentation has not correlated well with other variables, nor with student achievement gains. Perhaps the problem is with its definition. Perhaps the problem is that mode of presentation really isn't an attribute of the content of instruction.

Who describes the content?

In most efforts to describe the enacted curriculum, teachers have been used to self report on their instruction. For *Reform-Up-Close*, independent observers from the research team also reported on selected days of instruction. Comparisons were made between observers' descriptions and those from teacher self reports. There was strong agreement between the teachers and observers (Smithson and Porter 1994).

From the perspective of policy research, teachers are probably the most important respondents, since it is teachers who make the ultimate decisions about what content gets taught to what students, when, and to what standards of achievement. Curriculum policies, if they are to have the intended effect, must influence teachers' content decisions. Since the period of instruction to be described is long (i.e. at least a semester) teachers and students are the only ones likely to be in the classroom for the full time. Since content changes from week to week if not day to day, a sampling approach such as would be necessary for observation or video simply won't work. While it is true that video and observation have been used to good effect in studying pedagogical practice, this has only worked well when those practices have been so typical that they occur virtually every instruction period. However, some pedagogical practices are not

sufficiently stable as to be well studied, even with a robust sampling approach (Shavelson 1981) such as the sampling methods used in the Third International Mathematics and Science Study (TIMSS).

Students could also be used as informants reporting on the content of their instruction. An advantage of using students is that they are less likely than teachers to report on intentions rather than actual instruction. A danger with using students as respondents is that students' ability to report on the content of their instruction may be confounded with their understanding of that instruction. For students struggling in a course, their reporting of instructional content might be incomplete and inaccurate due to their own misunderstandings and lack of recall. We conclude that it is more useful to look to teachers for an accounting of what was taught, and to students for an accounting of what was learned.

Response Metric

When having respondents describe the content of instruction, not only must the distinctions in type of content be accurately presented as discussed above, but respondents need an appropriate metric for reporting the amount of emphasis placed on each content alternative. The ideal metric for emphasis is time; how many instructional minutes were allocated to a particular type of content? This is a metric that facilitates comparisons across classrooms, types of courses, and types of student bodies being served. But reporting number of instructional minutes allocated to a particular type of content over an instructional year is not an easy task. Other response metrics include number of hours per week (in a typical week), number of instructional periods, how frequently the content is taught (e.g., every day, every week) and percent of instructional time per year or per semester. The issue is how to get a response metric

as close to the ideal as possible and still have a task which respondents find manageable and which they can use with accuracy.

How frequently should data be collected?

There is a tension between requiring frequent descriptions to get accuracy in reporting (which is expensive) versus less frequent descriptions covering longer periods of instruction (say, a semester or full school year), which is less expensive and less burdensome, but may be less accurate as well. The issue is what frequency of reporting has an acceptable cost and still provides acceptable accuracy. We have used daily logs, weekly surveys, surveys twice a year and a single survey at the end of the year. When comparing daily logs to a single end-of-year survey, the results were surprisingly consistent (Porter, et.al. 1993).

In addition to cost and teacher burden, determining the instructional unit of time that should be described could also affect decisions about the frequency of reporting. At the high school level, the unit might be a course, but some courses are two semesters long while others meet for only a single semester. Alternatively, the unit might be a sequence of courses to determine, for example, what types of science a student studies when completing a three-year sequence of science courses. At the elementary school level policy makers are typically interested in a school year or a student's entire elementary school experience (or at least the instruction experienced up to the state's first assessment).

Summary and Conclusions

There are a number of important uses to be made of good quantitative information about the content of the enacted curriculum. There are, however, a number of issues in defining such

indicators: grain size, language, response metric, period of time to be described, and appropriate respondent. Several illustrations were provided of past efforts to measure the content of the enacted curriculum. Examples came primarily from high school mathematics and science, though similar work has been done at the elementary and middle school levels. The primary method used to collect the information is teacher self-report using survey instruments, including daily class logs, though observations have been used as well.

While teacher log data has been shown to agree quite well with accounts from observations by researchers and teacher questionnaire data has been a good predictor of teacher log data (Smithson and Porter 1994), perhaps the best indicator of the quality of survey data comes from using teacher self-reports to predict gains in student achievement. The upgrading mathematics project (Gamoran, et.al. 1997) reported the correlation at the class level to be about .5 and at the student level about .25.

The language given to teachers for describing the content of their instruction can also be used for content analyses of standards and tests, and alignment between the enacted curriculum and tests or standards can be determined. This was illustrated by content analyzing state tests, as well as the National Assessment of Education Progress and by determining alignment between teacher self-reports of their instruction and tested content. A great deal of attention is being given to alignment in standards-based reform, and these methodologies appear to be useful tools for determining where alignment exists and where it does not.

Teachers have also found the instruments to be useful in helping them to reflect upon their practice. In that sense then, teachers have provided validation for decisions made about grain size, language, response metric, and period of time to be described. Finally, as a standards-

based approach to reform continues to be utilized by states as a key feature of educational improvement efforts, interest in and need for useful descriptions of practice, assessments and standards will become increasingly important for answering questions about the implementation of standards in the classroom.

References

American Association for the Advancement of Science. 1989. *Science for all Americans. A Project 2061 report on literacy goals in science, mathematics, and technology.* Washington, DC: American Association for the Advancement of Science.

Blank, Rolf K., Jason J. Kim, and John L. Smithson. 2000. *Survey results of urban school classroom practices in mathematics and science: 1999 report.* Norwood, MA: Systemic Research.

Blank, Rolf K., and Doreen Langesen. 1999. *State indicators of science and mathematics education 1999. State-by-state trends and new indicators from the 1997–98 school year.* Washington, DC: Council of Chief State School Officers, State Education Assessment Center.

Bryk, Anthony S., and Steven W. Raudenbush. 1992. *Hierarchical linear models: Application and data analysis methods.* Newbury Park, CA: Sage Publications.

Council of Chief State School Officers. 2000. *Using data on enacted curriculum in mathematics and science: Sample results from a study of classroom practices and subject content.* Washington, DC: Council of Chief State School Officers.

Gamoran, Adam, Andrew C. Porter, John L. Smithson, and Paula A. White. 1997. Upgrading high school mathematics instruction: Improving learning opportunities for low-achieving, low-income youth. *Educational Evaluation and Policy Analysis* 19 (4): 325–38.

Garet, Michael S., Beatrice F. Birman, Andrew C. Porter, Laura Desimone, and Rebecca Herman (with Kwang Suk Yoon). 1999. *Designing effective professional development: Lessons from the Eisenhower Program.* Washington, DC: U.S. Department of Education.

Kennedy, Mary. 1999. Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis* 21 (4): 345–63.

Meyer, Robert H. (1997). Value-Added Indicators of School Performance: A Primer. *Economics of Education Review;* v16 n3 p283-301 June 1997.

National Council of Teachers of Mathematics. 2000. *Principles and standards for school mathematics.* Reston, VA: National Council of Teachers of Mathematics.

National Research Council. 1996. *National Science Education Standards*. National Academy of Sciences. Washington, DC: National Academy Press.

Peak, Lois. 1996. *Pursuing excellence: A study of U.S. eighth-grade mathematics and science teaching, learning, curriculum, and achievement in international context: Initial findings from the Third International Mathematics and Science Study* (NCES 97-198). Washington, DC: National Center for Education Statistics.

Porter, Andrew C. 1991. Creating a system of school process indicators. *Educational Evaluation and Policy Analysis* 13 (1): 13–29.

Porter, Andrew C. 1998. Curriculum reform and measuring what is taught: Measuring the quality of education processes. Paper presented at the annual meeting of the Association for Public Policy Analysis and Management, 31October, New York City.

Porter, Andrew C., Robert Floden, Donald Freeman, William Schmidt, and John Schwille. 1988. Content determinants in elementary school mathematics. In *Perspectives on research on effective mathematics teaching,* ed. Douglas A. Grouws and Thomas J. Cooney, 96–113. Hillsdale, NJ: Erlbaum.

Porter, Andrew C., Michael S. Garet, Laura Desimone, Kwang Suk Yoon, and Beatrice F. Birman. 2000. *Does professional development change teachers' instruction? Results from a three-year study of the effects of Eisenhower and other professional development on teaching practice* (Final report to the U.S. Department of Education on Contract No. EA97001001 with the American Institutes for Research). Washington, DC: American Institutes for Research.

Porter, Andrew C., Michael W. Kirst, Eric J. Osthoff, John L. Smithson, and Steven A. Schneider. 1993. *Reform up close: An analysis of high school mathematics and science classrooms* (Final report to the National Science Foundation on Grant No. SAP-8953446 to the Consortium for Policy Research in Education). Madison: University of Wisconsin–Madison, Consortium for Policy Research in Education.

Rowan, Brian. 1999. Assessing teacher quality: Insights from school effectiveness research.

Schmidt, William H., Curtis C. McKnight, Leland S. Cogan, Pamela M. Jakwerth, and Richard T. Houang. 1999. *Facing the consequences: Using TIMSS for a closer look at U.S. mathematics and science education.* Boston: Kluwer.

Shavelson, Richard J., and Paula Stern. 1981. Research on teachers' pedagogical thoughts, judgments, decisions, and behavior. *Review of Educational Research* 51 (winter): 455–98.

Smithson, John L., and Andrew C. Porter. 1994. *Measuring classroom practice: Lessons learned from efforts to describe the enacted curriculum—the Reform-Up-Close study* (CPRE Research Report Series #31). Madison: University of Wisconsin–Madison, Consortium for Policy Research in Education.

Wang, Jia 1998. Opportunity to Learn: The Impacts and Policy Implications. *Educational Evaluation and Policy Analysis.* v20 n3 p137-156 Fall 1998.

Ware, Melva, Lloyd Richardson, and Jason Kim. 2000. *What matters in urban school reform.* Norwood, MA: Systemic Research.

Webb, Norman L. 1997. *Criteria for alignment of expectations and assessments in mathematics and science education* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 6). Madison: University of Wisconsin–Madison, National Institute for Science Education.

Webb, Norman L. 1999. *Alignment of science and mathematics standards and assessments in four states* (Council of Chief State School Officers and National Institute for Science Education Research Monograph No. 18). Madison: University of Wisconsin–Madison, National Institute for Science Education.

Table 1
AVERAGE ALIGNMENT
TEST to TEST

	State to State	NAEP to State
Math 4	0.41	0.35
Math 8	0.33	0.30
Science 4	0.33	0.29
Science 8	0.28	0.20

Table 2
AVERAGE ALIGNMENT
INSTRUCTION to TEST

	Target State	Other States	NAEP
Math 4	0.42	0.33	0.41
Math 8	0.33	0.24	0.22
Science 4	0.37	0.28	0.23
Science 8	0.33	0.23	0.14

Table 3
AVERAGE ALIGNMENT
INSTRUCTION to INSTRUCTION

Math 4	0.8
Math 8	0.68
Science 4	0.7
Science 8	0.64

Table 4
AVERAGE INTER-RATER AGREEMENT
ON ASSESSMENT ANALYSES

	Fine Grain	Med. Grain	Lrg. Grain
Elementary Math	.47	--	.70
Distinctions Possible	*(438)*	*(N/A)*	*(36)*
Middle Sch. Math	.47	--	.70
Distinctions Possible	*(504)*	*(N/A)*	*(36)*
Elementary Science	.40	.50	.56
Distinctions Possible	*(396)*	*(84)*	*(30)*
Middle Sch. Science	.38	.56	*not available*
Distinctions Possible	*(876)*	*(150)*	*(36)*

Table 5
Indicators of Instructional Alignment Correlations with Achievement Gains

Multiple regression using Level and Configuration	*r* *Class Gains*	*r* *Student Gains*
Topics Only	.260	.245
Cognitive Demand Only	.106	.166
Topics by Cognitive Demand	.481	.260

Table 6
Middle School Science
Categories of Cognitive Demand
Evaluation of Eisenhower Program Longitudinal Study

The following list identifies key discriptors for each category of cognitive demand. Refer to this list in considering your responses for each category of cognitive demand on those topics covered as part of science instruction.

Memorize
- Facts
- Definitions
- Formulas

Understand Concepts
- Explain concepts
- Observe teacher demonstrations
- Explain procedures/methods of science & inquiry
- Develop schema, or frameworks of understanding

Perform Procedures
- Use numbers in science
- Do computation, execute procedures or algorithms
- Replicate (illustrative or verification) experiments
- Follow procedures/instructions

Generate Questions/Hypotheses
- Brainstorm
- Design experiments
- Solve novel/non-routine problems

Collect Data
- Make Observations
- Take measurements

Analyze & Interpret Information
- Classify/order/compare data
- Analyze dat, recognize patterns
- Infer from data, predict
- Explain findings, results
- Organize & display data in tables, graphs, or charts

Use Information to Make Connections
- Use & integrate concepts
- Apply to real-world situations
- Build/revise theory
- Make generalizations