Practical Assessment, Research, and Evaluation

Volume 9 *Volume 9, 2004* Article 14

2004

Designing Accountability Assessments for Teaching

William D. Schafer

Mark Moody

Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 14, July, 2004 ISSN=1531-7714

Designing Accountability Assessments for Teaching

William D. Schafer, University of Maryland & Mark Moody, Maryland State Department of Education

We argue (1) that the domains of externally mandated tests should be transparent to teachers, (2) that test maps can be used to communicate with teachers, and (3) that the most effective test maps express domains in terms of combinations of what students know and can do. We propose a device called a heuristic as the lowest level of specificity in a test map, and criteria to achieve a balance between too much specificity and too much generality in heuristics. Although we focus primarily on summative assessments, we also discuss how a state may use related, teacher-developed formative assessments to foster information-rich classroom environments. Other than item bank structure and sampling, our suggestions would have little impact on how tests are developed or scored, but we believe they would have a large impact on how they are described and used, and ultimately on their effectiveness as agents of instructional reform.

Need for Domain Description

As Baker (2002) recently noted, defining "the operational limits of the target domain of learning" is a necessary condition for using assessments effectively for both accountability and for school improvement. But if assessments are to direct reform, the achievement targets that constitute the domain of each of these tests must (a) be a legitimate domain of achievement targets (by this, we mean that agreement has been reached using an accepted process), (b) be sufficiently described to be communicated effectively to others, especially instructional personnel, and (c) be reliably sampled by the test (i.e., not only does the test sample the domain well, but also, teachers believe it will sample the domain well). These principles apply to any test originating from outside the classroom and intended to effect change in the classroom. Each is discussed briefly.

(a) The test must cover a legitimate domain.

Like most people, educators work more effectively if they believe their goals are worthy. In education, that means the value of the targets of instruction is apparent. While each of us can and often do make judgments according to our own beliefs about any set of curricular goals (i.e., learning targets), harnessing the efforts of schools, districts, and an entire state requires a shared belief in the worth of the goal. In our democratic society, that requires a process, usually political, in order to attain a consensus. For example, in developing goals in some content area, a process that effectively includes representation of teachers will be better accepted by educators than one that does not. The state school board, representing a broad constituency of stakeholders, can be an accepted authority to approve both the process and the product.

As Popham (2004) has noted, curricular goals are often too broad to be covered completely by teachers. There typically are too many curricular goals either to cover comprehensively or to be represented completely in an assessment program. Popham concluded that curricular goals must be reduced, both in order to focus teaching on a manageable set of learning targets that are of most importance, and in order to allow valid assessment of student achievement across that domain.

Popham (2004) suggested three approaches to reduce the domain of instructional targets: new standards, coalesced standards, or derivative assessment frameworks. New content standards would be time-consuming to develop. By coalesced standards, Popham described a hierarchical structure in which current targets are subsumed under fewer, but nevertheless measurable goals, but that, too, would be time-consuming and perhaps difficult to achieve. Given these drawbacks, Popham concluded that the third option, derivative assessment frameworks, is most likely to succeed. That is the approach taken here.

(b) The domain must be effectively described.

A test is supposed to assess what students are to know and are able to do with what they know. Actually, though, every achievement test item prompts both these elements because it requires a student to do something with something. Classroom assessment textbooks such as Nitko (2001) almost universally recommend a table of specifications as a device for describing test items in terms of the content and process dimensions. That is, what a student is expected to know and

what he or she is expected to do with that knowledge is described by combinations of content (e.g., rows) and process (e.g., columns) in a table of specifications.

But a table of specifications seems inadequate to communicate the domain of a test to those in the field who need to understand it in terms of the instructional targets it represents. Both dimensions of the table are too imprecise and the process dimension may not apply to the content elements in the same way to different educators, including teachers and test developers.

The content dimension typically is not sufficiently detailed to determine the extent of that which students must know. The ambiguity is tolerated, perhaps for several reasons. These may include avoiding limitations on the scope of what teachers may teach, making sure that teachers do not teach to the test, or making the table convenient to express. But each of these is not an appropriate goal for a table of specifications. The domain of a test does not, in and of itself, limit what teachers may teach, although it does describe what the agency that mandates the test expects teachers to include in their instruction. Teaching to a test domain is not teaching to a test; teaching to a test should be used to describe only limiting instruction to a biased sample of the domain, such as those items on a particular form of the test. A table of specifications should be judged on whether it affects practice, not on whether it is easy to describe. Elaborations of the content elements are necessary in order that the assessment specialists who write tests and the educational specialists who use them agree on the scope of the knowledge elements.

Similarly, the process dimension requires clarification. It is generally acknowledged as inadequate among assessment professionals that students be asked only to recall content knowledge, but specifying the higher-order reasoning that is to be included in an instructional domain is not straightforward. Likely, different educators would disagree on even the way higher-order thinking should be described. For example, Nitko (2001) describes four approaches in his introductory classroom assessment text. Nevertheless, a complete domain description should indicate not only what students are to know, but also what they are expected to be able to do. Otherwise, educators (especially curriculum developers and teachers) and assessors will not be working toward the same domain.

Even a highly motivated educator cannot attain a goal that is unclear. Some way is needed to clarify the domain of each test so it can communicate unambiguous targets in combinations of both content and process dimensions. We will suggest below a way to clarify both dimensions.

(c) The domain must be reliably sampled.

Everyone agrees that the domain of any assessment should be sampled representatively on each test form. However, teachers who have worked with mandated assessments often do not feel the test covers what they have been teaching, even when they honestly believe they have represented their district's curriculum instructionally. Perhaps they are often right. The connection between the tested domain and the educators' learning targets needs to be established at the start of the appropriate instructional sequence. Not only must the educator understand the domain, but he or she must also believe the test will sample it appropriately. Otherwise, the test will be marginalized as irrelevant and teacher motivation expected as a result of the assessment and accountability program could be lost.

In summary, a testing program is effective as a guide to instructional goals to the extent that it covers a publicly accepted learning domain that is described in terms of both content and cognition.

Test Maps

A test map, which is usually more specific than a table of specifications, describes the content of the test and how it is sampled to produce each form. Examples of test maps may be found at the web page http://mdk12.org/mspp/high_school/look_like/index.html where there are several sample tests that follow test maps for a high school assessment program. We propose a modification so that a test map may be used as a unifying device to express a legitimate achievement domain in unambiguous terms and to ensure not only that any form of the test will sample that domain appropriately, but also that educators will anticipate appropriate coverage (conforming to our three principles, above). Because we know it best, we use Maryland's assessment and accountability program as our illustration. We point out where Maryland's program illustrates our recommendations, but the majority of our suggestions have not to our knowledge been implemented anywhere and would apply just as well to any other mandated assessment program as they would to Maryland's.

Our recommendations apply to any assessment program in which teachers are motivated by their students' test performance. Normally, that means the program includes consequences (rewards or sanctions), either for students or for providers of instruction. The stronger those sanctions, the stronger the motivation they presumably provide.

Assessments written to a test map developed by a state need not be the only assessments taken by students, of course. In addition to state assessments, districts may write augmentations to expand the curricular goals for teachers and for test developers beyond those at the state level. Indeed, we could envision schools and teachers doing the same. But at each level, there is a need for clarity of goals. Thus, the process of test map development would be equally helpful in these broadened assessment programs as they are at the state level.

We first describe current practices using Maryland as an example. We also describe some shortcomings that we feel limit the potential of our current assessments to direct instruction. Then we discuss how existing materials may be focused into derivative assessment frameworks, as recommended by Popham (2004), called test maps. Our test maps use what we call "heuristics," designed to express both activity and content dimensions of achievement targets in a way that defines instructional targets for teachers and focuses assessment developers on those targets, but does not threaten test security.

Example of Current Practice

We use an example from the Maryland's State Content Standards to illustrate some of the aspects of our proposal. The full State Content Standards may be found at http://mdk12.org/mspp/standards/index.html and from there, links may be followed to any of the cited material that follows.

The State Content Standards describe the expected domain of education for students in the state in four content areas: Language Arts, Mathematics, Science, and Social Studies. Our example will be taken from Mathematics.

Within each content area, there are several expectations. We will use the expectation level as the degree of specificity of content in our example. In Mathematics, there are ten expectations: (1) Knowledge of Algebra, Patterns, and Functions, (2) Knowledge of Geometry, (3) Knowledge of Measurement, (4) Knowledge of Statistics, (5) Knowledge of Probability, (6) Knowledge of Number Relationships and Computation, (7) Process of Problem Solving, (8) Process of Communication, (9) Process of Reasoning, (10) Process of Connections. We will use Knowledge of Probability in our example.

Within each expectation there are several indicators. The indicator is the lowest level of specificity in the statement of the domain that is to be represented by Maryland's tests. In probability at the eighth-grade level, there are five indicators that build upon the three indicators at the fifth-grade level, which in turn build upon the three indicators at the third-grade level. This sort of representation of an instructional domain is likely not unusual. Assessments are commonly constructed using indicators.

Among the eighth-grade extensions is the indicator "find the probability of simple dependent and independent events using various methods including constructing a sample space." This statement is typical of the level of detail in most state descriptions of learning targets. But it is our contention that an indicator like this is inadequate to describe for a teacher what needs to be covered during instruction.

Should someone disagree and feel the statement is adequate, let's say you are asked to teach students to "find the probability of simple dependent and independent events." What do you include? Do you express probability as a ratio of equally likely events, as the limit of repeated samples, as a degree of belief, or as some combination of these? Do you define simple and compound events? Do you express dependence as limiting the sample space to a subpopulation or do you define unions, intersections, negations and dependence vs. independence and then use computational formulas for probabilities? Do you teach Venn diagrams? Do you teach your students to use two-way arrays for computing probabilities of conditional events? These questions are important; they speak directly to what students will be asked to do in the classroom and on the test. The answers to these questions are crucial for alignment to exist between instruction and assessment. But how should you as a teacher answer questions such as these when you are told only to teach students to "find the probability of simple dependent and independent events"?

One approach to your dilemma would be to guess. Indeed, what else can you do? But there is no guarantee that the test will cover the domain the way that you decided to teach it. Of course, you should look at your district's curriculum materials. But even if they answer these questions, there is no guarantee that they cover the same domain as the test will since educators who themselves were guessing about these and other questions like these wrote them. All they do for consistency is help you make some of the same guesses as do the other teachers in the district, and these will almost certainly vary district-to-district.

Parenthetically, it is sometimes mentioned that some ambiguity is helpful, since it encourages teachers to teach a broader array of material. Certainly, some may do that. Others may decide to concentrate on one, but not all ways to cover a particular indicator. There may be other solutions to get from an ambiguous indicator to an individual teacher's instructional goals. But that seems a rather haphazard approach to curriculum. We can easily imagine that some students will not have had an opportunity to learn material that will be on the state test due to a misunderstood content domain. Others may have been instructed in a domain that stretches beyond the scope of the test's content domain, so that the test under-samples their achievements. While allowing a district, a school, or a teacher to enhance the scope of its curriculum beyond the objectives of the state is to be encouraged, to do so through planned ambiguity in mandated tests seems a poor policy. To the extent that the domain of the test is ambiguous, students' opportunity to learn the tested domain becomes haphazard.

Assessment Limits

Since the state is the authority that is entrusted with accountability testing, it is the state's responsibility to ensure that teachers understand the scope of the indicators that are to be taught and that the domain of the test agrees exactly with

the extent of these limits. Without these two requirements, we cannot make inferences about the causes of poor test scores, nor will we be able to do much to improve them. Just as learning targets need to be clarified for students and then assessed as they have been clarified, so also is it necessary for teachers to understand their instructional targets in order to hit them. Hence our title; our purpose is to support teaching by using assessments to clarify learning domains.

Any test is an operational definition of some domain and was constructed according to rules that included and excluded certain specific content elements. While these rules must exist, unless they are written down, they exist only in the mind(s) of those responsible (at the time the test is constructed) for the test's content. Thus, they are hidden from teachers.

Let us explore how indicators may be made more explicit. We will introduce and then elaborate upon the concept of Assessment Limits (a term borrowed from Maryland's assessment programs; see the Appendix for Maryland's statement about assessment limits as they are used in its biology testing program). Assessment limits as presently used specify the exact content that may appear on the test. When developed properly, they define what is and is not "fair game" for the assessment. They represent statewide consensuses that were developed with broad teacher representation. Assessment Limits are widely disseminated and are used in item and test development.

Here is an example that we constructed of Assessment Limits for our example indicator. A few of the limits actually apply also to other indicators in Maryland's eighth-grade mathematics expectations, but are included anyway so they appear more internally coordinated.

1. A universe is the entire collection of outcomes that may occur.
2. An event is an occurrence or outcome that satisfies a condition.
3. Mutually exclusive events may not occur together in one outcome.
4. The condition that defines an event may be simple (based on one characteristic) or compound (based on two or more characteristics).
5. Compound events (or conditions) are derived from simple events by parentheses and the operators union ($\cup$), intersection ($\cap$), and negation ($\sim$). The assessment is limited to at most three events, two unions, two intersections and one negation in any compound event.
6. Probability of an event, P(A), is the frequency of occurrences that satisfy the event (A) over the total frequency of occurrences. This definition requires an assumption that all occurrences are equally likely.
7. Probability of an event is also the limit of the ratio of number of times the event occurred over the number of trials as number of trials increases. This definition requires an assumption that the trials are independent.
8. Express a universe of two sets of mutually exclusive events as a two-way table of frequencies or probabilities. Annotate the table to show derivable compound events in the cells and in the margins.
9. Express a universe of three sets of mutually exclusive pairs of events as a Venn diagram. Annotate the diagram to show derivable compound events for all regions.
10. The relation "given" (|) limits the universe to outcomes satisfying the event following the relation.
11. Two events are independent if the probability of one is unaffected by the (non)occurrence of the other. That is A and B are independent if $P(A \mid B) = P(A)$.
12. $P(A \cap B) = P(A) \cdot P(B \mid A)$.
13. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
14. $P(A \mid B) = P(A \cap B) / P(B)$.

While these limits are painstaking to enumerate, they communicate an unambiguous sub-domain to a teacher (and to an item writer). Of course, these are just our example. They do not represent the position of Maryland on the meaning of that indicator. But the mathematics education community could easily reach consensus on a similar list for this and all other indicators. These would then become at once an "at least list" for teachers and an "at most list" for test developers who are writing items for that indicator to represent the core knowledge embodied in the indicator. The list defines the content of test items that may be sampled to represent the content of the indicator. The sampling process also must ensure that no element on the list has a zero probability of appearing on the overall assessment. Otherwise, the assessment limits would not describe the realized assessments.

Note also that this sample of assessment limits defines a sub-domain that is instructionally meaningful to assess. If precision were adequate for interpretation, a sub-score would carry implications for individual student remediation as well as for modifying an instructional program. Some others of the indicators for the probability expectation might be amalgamated into this sub-domain and remain instructionally meaningful, as well. Indeed, it would be valuable for a state or other appropriate education unit to organize each of its content domains around instructionally valuable sub-domains and thus to be able to generate interpretable sub-scores from them.

Extended Assessment Limits and Heuristics

Identifying content assessment limits only gets us part of the way toward being able to communicate achievement targets to teachers and stakeholders in terms of what students need to know and be able to do. It is also necessary to describe the cognitive activities that constitute the behavioral part of each student expectation. We describe a possible way to include cognition in a test map and thus to meet the two criteria (domain understanding and assessment

representation) we believe are necessary (though certainly not sufficient) for effective tests.

One of the problems we face is that there is no accepted codification of cognition. The Bloom, Englehart, Furst, Hill, & Krathwohl (1956) taxonomy is best known among educators, but it has been reported to be difficult to use, does not correspond to valued cognition outcomes such as problem solving, and is not universally taught in teacher preparation. That it has been updated recently (Anderson & Krathwohl, 2000) may help make it more useful, but at the same time contributes to a lack of consistency. Among those referenced by Nitko (2001), Marzano, Pickering & McTighe (1993) seems to be most useful for our purposes. They identify thirteen reasoning strategies that seem to be more consistent with publicly valued outcomes (e.g., one of them is problem solving). As Nitko (2001) points out, each of these outcomes has implications for assessment.

Another problem is that the descriptive language of cognition is not consistent across disciplines. Each discipline has a unique perspective nationally. If cognition were to be described equivalently across subject matter areas, then we think teachers would naturally become more focused on teaching cognition (and likely metacognition) in order to teach efficiently, thus drawing useful parallels among disciplines. However, the structure and language of our current domain specifications at the national level are not at all equivalent in their representation of cognition, in part because we lack a nationally accepted and used taxonomy. This leaves teachers with a belief that cognition is discipline-specific, probably discouraging applications across contents and even attention to higher-order thinking, in general. It is likely naïve to expect any state to define any content domain in a way that is dramatically different from its national parallel(s), so the obvious approach to representing cognition by using some codification of thinking and applying it to all content standard descriptions is unlikely.

Rather, we suggest application of a state-endorsed codification at the level of content assessment limits. Our suggestion is to ask the same state-level content experts who agreed on the content assessment limits to recommend, for each limit, the elements in the accepted taxonomy for defining the cognition assessment limits that will be "fair game" for the test. Their deliberations should be conducted with the input of experts in cognitive processes. For example, say we used the Marzano et al. (1993) taxonomy with the "finding probability" indicator's content limits as described above. Then we would take each limit and ask which cognitive categories will be assessed. For the first content limit, which is "a universe is the entire collection of outcomes that may occur," it seems to us to make sense to ask students to these four out of the thirteen reasoning strategies:

1. generalize a statement of a universe from a description of its elements (the taxonomy category name is induction),
2. identify whether new elements are or are not members of a given universe (the taxonomy category name is deduction),
3. correct a misstatement of universe (the taxonomy category name is error analysis), and
4. explain why a given statement of a universe is adequate for a given purpose (the taxonomy category name is constructing support).

We now have four statements that describe specific ways to use the definition of a universe that can guide both teachers and test developers. Note that the statements are at the appropriate level of generalization for what have elsewhere been called "heuristics" (Schafer, 2002). A heuristic must be specific enough for making judgments about whether test prompts measure them, but general enough so that there are virtually an unlimited number of such prompts that could be written. These criteria are borrowed from Kerlinger (1990), who argued that an effective conceptual definition of a construct should be general enough to allow multiple operational definitions, but specific enough that the validity of any given operational definition will be apparent. A heuristic describes what a student is to know and do.

How many heuristics would there be for an expectation? In the Maryland eighth-grade mathematics standards for the probability expectation, there are five indicators (see above). We identified fourteen content assessment limits, but some of these represent other indicators, there is some degree of overlap among the indicators. It seems reasonable to assume that there might be five unique content assessment limits for an average indicator. We found four cognition assessment limits to apply to the first content assessment limit. Using that as a typical number, we estimate that a total of 5 x 5 x 4 = 100 heuristics might apply to a typical expectation. This seems manageable to us as a domain for an assessment, but if a state-level team of teacher content experts feel it is too ambitious for instruction, then part of their task should be, by consensus, to pare the list down to its essentials, making sure to include all appropriate factual, conceptual, and procedural knowledge (Anderson & Krathwohl, 2000). The intent is to describe the appropriate educational domain for instruction. The state-determined assessment will then appropriately and openly represent the agreed-upon learning targets. Decisions at this level about legitimate limits are made for every test we now use; they have to be to get from indicators to items. Yet the decisions are not only hidden, but depend on who happens to be constructing the test at the time. We are really arguing here to make explicit (and consistent) something that is actually being done in every testing program.

For a complete domain description, the heuristics may be augmented further with assessment examples to help communicate their precise intent. For example, for the induction example, one could ask a student to display a universe of simple outcomes when two eight-sided dice are to be thrown. The items should correspond to the types of items that will be used on the state's summative assessments.

We are recommending the use of test maps based on heuristics to define agreed-upon instructional targets. The test

maps thus created define the scope of what the state values in each tested content-grade-level combination. In order to guide curriculum and instruction, they should be freely available throughout the state. The maps should also indicate the rules that are to be followed to sample the heuristics that are to be represented on forms of the test (but not, of course, the result of those rules for a particular test form that is to be used). That way, alignment between mandated domains of learning targets and mandated assessments is not a matter of research; it is a guarantee. Further, curriculum specialists and teachers can use the domains to align instruction to them.

Summative and Formative Assessments

A state's summative assessments could be administered separately over the instructionally valuable sub-domains (e.g., expectations) discussed earlier. Indeed, students might even pass (or fail) contents based upon scores earned over assessments of sub-domains (e.g., expectations such as "probability"), which might be administered when their teachers deem students as ready for them. Instead, all sub-domains are typically taken all at once during a testing window, such as at the end of the year.

Computers can play a useful role in individualizing assessments in order to make on-demand assessment at the sub-domain level feasible and to control item overexposure. If overall content scores are needed (e.g., to generate a proficiency level result for accountability purposes), the sub-domain scores could be aggregated up for that purpose. On-demand, individualized summative sub-domain assessments could be administered by school testing coordinators at secure sites in schools.

Companion formative assessments should be available so that teachers have the resources to judge readiness in "real time," according to each teacher's schedule of instructional decisions that need to be made. Students and teachers need to know on a day-to-day basis where they are with respect to achieving proficiency on the content standards and indicators. Teachers need the information to adjust their strategies to address each student's needs. Students need the information to guide their learning. The formative assessments may or may not be curriculum-embedded. The state is the appropriate agency to develop these since it is the "owner" of the test maps and the proficiency standards. Teachers could play a valuable role in developing and disseminating their own materials through state-administered channels.

The model used in academic fields to review and disseminate scholarly works could work here. For example, a state might establish a process by which teacher-developed formative assessments that correspond to test maps at the sub-domain level could be forwarded to a refereeing board. Developmental work could even be supported by the state through solicited or unsolicited grants to individual teachers or to teacher groups. Teacher-developed formative assessment submissions might require some data to document effectiveness; perhaps, for example, involving an independent tryout of the assessments in other classrooms. The board might review the submissions by a standard process, perhaps making recommendations to the author(s), and accept (or reject) the resulting documented formative assessments for dissemination throughout the state. Accepted assessments then could be made available throughout the state using a searchable database. The entire process could be accomplished in-house or through a vendor.

Of course, there should be some incentives for teachers to produce acceptable formative assessments. In the academic world, typical rewards are recognition (prestige), salary, and promotion. Teachers might receive similar rewards for successful productivity (through acceptance and dissemination) of formative assessments. The teacher-author's school and district could appear along with his or her name(s) to provide institutional incentive for productivity, much as is done in a university. An interesting by-product could be an increase in the professionalism of teaching and perhaps enhanced job satisfaction of successful teachers.

Conclusion

If we are to see any substantive improvements in student achievement as a result of assessment and accountability, we must be able to have a significant impact in the classroom. After more than 10 years of assessment driven reform at the state level in Maryland (our example), we believe perhaps even a majority of the state's teachers do not yet understand what proficient student work looks like in the same way it is understood at the state level. We have noticed that virtually everyone who observes classrooms seems to come away with a similar conclusion.

But fault does not lie with teachers. Educators at the state level have not established a meaningful link between content standards and day-to-day student performance. We are convinced that Maryland is not alone. Most if not virtually all other states have also failed "unpack" their standards and indicators so that they are understandable and explicit as guides for classroom instruction. Rethinking the way we deliver summative and formative assessments is a logical extension of this argument, one that treats assessments as instructional tools.

The critical first step in integrating meaningful assessment strategies into day-to-day instruction is full articulation of the content and cognition of learning domains. Teachers and students must be able to understand what the essential characteristics of a proficient response are at the indicator level. Classroom activities should be created that provide students with opportunities to demonstrate the depth of their understandings and that also provide teachers with a rich source of diagnostic information to help them understand each student's strengths and needs with respect to attaining proficiency. The teachers focus in evaluating student work shifts from sorting students based on their performance to [illegible]. We believe it would best serve its education community for a state to

represent its augmented assessment limits, both the statements of content and their elaborations (heuristics), in a test map that will serve to explain to teachers exactly what the target is in terms of the characteristics of a proficient response. Not only does this approach remove ambiguity about what is fair game for the test, but also it will serve to guide test developers as they create new editions, or forms, of the state's assessments. The map should enumerate all the content and cognition limits, should describe what formats will be used to represent them, should specify how and with what frequency, or probability, they will be sampled, and should explain what sub-scores will be developed (and how they are developed) from the student responses. Since no unreleased test items are needed, the map would not violate test security and could (we believe it should) be freely available throughout the state.

While this seems like a tall order, a state has the resources and, we argue, the responsibility to achieve a consensus among stakeholders around the content to be assessed. Whether or not they are articulated or even thought consciously about, decisions at the level of detail we describe must be made to build any test since every test item must ask a student to do something with something. Every test represents someone's understandings about the domain and its limits. By developing and using an explicit map, a healthy consensus about what should be taught will be reached, teachers and other stakeholders will no longer be guessing about test content, test developers will produce assessments that represent appropriate domains, sub-scores will represent meaningful information for both students and programs, and there will be no surprises in the assessment system. A state-developed system of articulated summative and formative assessments, representing explicit test maps expressed in terms of heuristics, can foster the clarity of achievement targets necessary for teachers and other educators to develop more efficient instructional activities and ultimately to effective and documented school reform.

In summary, we believe that failure of assessment-driven reform is at the state level because there is no visible, established link between assessed student performance and day-to-day instruction. To remedy this, a state can develop for its educators complete descriptions of what students are expected to know and be able to do in order to provide clarity of learning targets, and summative and formative assessments to provide needed tools. We have tried to describe some steps a state could take to accomplish these goals.

Note:

This project was partially funded by the Maryland State Department of Education (MSDE) under a contract to the Maryland Assessment Research Center for Education Success (MARCES). The opinions expressed do not necessarily reflect those of MSDE or of MARCES.

Appendix

"Understanding Assessment Limits" in Biology from the Maryland Web Site

The Science Core Learning Goal (CLG) document is used for both assessment and instruction. The "assessment limits" included in the 1999 document were derived from the "at least" list that was included in the 1996 document. Assessment limits help clarify what a student will be asked to know, what a teacher will be asked to teach, and the content from which test questions will be drawn. The Maryland State Board of Education (MSBE) requires that all students have the opportunity to learn content about which they will be assessed. The clarification of content in the assessment limits supports this requirement.

Assessment limits can be thought of in two ways: for *instruction*, they represent the **minimum** content that must be taught (the course must include **at least** the content outlined by the assessment limits); for *assessment*, they represent the **maximum** domain from which test questions will be developed (assessment limits identify the content which is fair game for the development of test items). All assessment items developed for the High School Assessments will be drawn from the assessment limits. However, not every assessment limit will be tested on every form of the test.

There are five science Core Learning Goals:

- Goal 1: Skills and Processes
- Goal 2: Concepts in Earth/Space Science
- Goal 3: Concepts in Biology
- Goal 4: Concepts in Chemistry
- Goal 5: Concepts in Physics

The skills and processes in Goal 1 are essential to science learning and will be assessed with each of the other four goals. In Goal 1, the indicators and the assessment limits **are identical**. Those marked "NT" will not be assessed on the biology test. However, they are still appropriate for instruction and other types of formative assessments.

The assessment limits included in Goal 3 (concepts of Biology) are a subset of the concepts that should be covered in a biology course. Goals 2, 4, and 5 do not include "assessment limits," per se. Since these content areas will not be assessed in Phase 1, they have not as yet been revised. Instead of assessment limits, these goals still contain an "at least" list. As

Maryland develops assessments for Goals 2, 4, and 5, their "at least" designation will also be changed to assessment limits.

An illustration of assessment limits follows. In the biology CLG, Expectation 3.3 deals with genetics. Indicator 1 states that, "The student will demonstrate that the sorting and recombination of genes during sexual reproduction has an effect on variation in offspring." The two assessment limits which follow indicator 3.3.1 state:

- meiosis (chromosome number reduced by one-half; crossing-over may occur)
- fertilization (combination of gametes).

Therefore, test questions derived from biology indicator 3.3.1 may include questions about how fertilization is related to variation in sexually reproducing organisms, specifically, the role of meiosis in producing gametes, in reducing chromosome number, and the inheritance of new traits that result from crossing-over. Test questions may *not* include items dealing with the steps of meiosis, the identification of structures present in cells during meiosis, or the structure of the organs or organ systems where meiosis occurs.

Vocabulary that is essential to understanding the concept being assessed may appear in an item, but vocabulary that relates to explicit details not essential to the understanding of an overall concept will not. For example, knowledge of trophic levels is critical to understanding food webs (3.5.4), but knowledge of Turner's Syndrome is not essential to understanding the effects of an abnormal number of chromosomes on an organism (3.3.4).

Some critics may say that the use of assessment limits means teachers will be "teaching to the test." However, the phrase "teaching to the test" is misleading and a misnomer. Obviously, one can not teach to a test since the test questions are not known. What teachers really do is teach to a target, the local school system curriculum, and devise appropriate assessments (tests) to check how well the students have learned what they were taught. The extent of student learning is assessed through observations, classroom quizzes, homework, written assignments, formal teacher made tests, structured laboratory activities, etc. How else will teachers know if their students have learned? The local school system curriculum should be closely aligned with the CLG, and formative assessments should prepare students for the end-of-course assessment.

Concern has been expressed that some teachers will adjust the curriculum to include only the content defined by the assessment limits. The "at least" portion of the original Core Learning Goals was designed to outline the non-negotiable content for a given course, not the entire course. Local principals, supervisors, and others must monitor instruction to insure that the curriculum being taught meets the requirements established by the local system. Reasonable requirements for coverage of the curriculum, pacing, grouping, and other instructional decisions are developed locally.

The 1999 Core Learning Goal documents also differ from previous versions through adjustments to a limited number of indicators and the removal of sample classroom learning activities. No changes were made in the goal or expectation statements, however, the language of certain indicators was modified if it was shown to be ambiguous or contained multiple actions for instruction and/or assessment. In cases of the latter, the actions were split between separate indicators. For example, an indicator that stated that students will *analyze* and *evaluate* was divided into separate indicators for each verb.

In conclusion, the CLG document represents the "core" content for both instruction and assessment. Local school systems should use it appropriately when making decisions about curriculum, instruction and assessment.

References

Anderson, L. W. & Krathwohl, D. R. (2000). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's Taxonomy of educational_objectives*. New York: Longman.

Baker, E. L. (2002). Visions of test results dance in their heads. *NCME newsletter*, *10*(2), 5-6.

Bloom, B. S., Englehart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. White Plains, NY: Longman.

Kerlinger, F. N. (1990). *Foundations of behavioral research* (3rd Ed.). Orlando, FL: Harcourt Brace.

Marzano, R. J., Pickering, D., & McTighe, J. (1993). *Assessing student outcomes: Performance assessment using the dimensions of learning model*. Alexandria, VA: Association for Supervision and Curriculum Development.

Nitko, A. J. (2001). *Educational assessment of students* (3rd Ed.). Upper Saddle River, NJ: Prentice-Hall.

Popham, W. J. (2004, June). *Changing curricular horses in mid-stream: Sometimes the river requires it*. Presented at the National Conference on Large-Scale Assessment, Boston, MA.

Schafer, W. D. (2002). How can assessment contribute to an educational utopia? In Lissitz, R. W. & Schafer, W. D. (Eds.) *Assessments in educational reform: Both means and ends*. Boston: Allyn & Bacon, pp. 80-98.

Descriptors: Academic Achievement; Curriculum Development; Educational Assessment; Educational Improvement; Feedback; Performance Based Assessment; Performance Factors; Standards; Test Construction; Test Use

Citation: Schafer, William D. & Mark Moody (2004). Designing accountability assessments for teaching. *Practical Assessment, Research & Evaluation*, 9(14). Available online: http://PAREonline.net/getvn.asp?v=9&n=14.