

User-Centered Assessment Design

Paul Nichols
David Mittelholtz
Jeremy Adams
Robert van Deusen

Pearson

Paper presented at the annual conference of the American Educational Research Association, New York, New York, March 2008.

In this paper, we introduce user-centered assessment design (UCAD), an approach to test design intended to produce assessments that deliver to teachers the kind of complex information on student learning and knowledge that they can combine with sound pedagogical practice to improve student achievement. First, we introduce UCAD and describe how test design might look under a UCAD approach. Second, we demonstrate UCAD using a pilot study of science assessment design for formative purposes. Finally, we conclude by comparing UCAD to a conventional approach to test development and to another innovative approach, evidence-centered test design (ECD; Mislevy, Steinberg, Almond & Lukas, 2006).

Introduction to User-Centered Assessment Design

The roots of UCAD are in user-centered design (UCD). In this section, we first review UCD. Then, we will introduce an approach to test design, UCAD, that incorporates UCD principles. We describe how test design might look under a UCAD approach.

Review of user-centered design

The term user-centered design (UCD) was popularized by Norman (1986). In the subsequent book, “The Psychology of Everyday Things,” Norman (1988) advanced a design philosophy that placed the needs of the product user at the center of product design. According to Norman, human problems and errors when dealing with technology are usually a result of design failure. Good design accounts for human capabilities and minimizes problems people encounter dealing with technology. He described a design process that focused on psychological variables linked to the user such as memory, strategies, mental models and learning. Finally, Norman introduced concepts and design principles that help to analyze bad design, to identify bad design elements, and to create good design.

Much of the subsequent operationalization of the UCD concept has been done in the field of human-computer interaction. Examples include web design (Hackos & Redish, 1998), PDA design (Lamberts, 2002) and music file sharing (Brown, Geelhoed, & Sellen, 2001). This development has resulted in a number of international standards that address different aspects of usability. Measurement procedures have been developed to assess product developers' success in meeting usability goals (see, for example, Bevan and Schoeffel, 2001).

The UCD applications have emphasized two features: collecting data directly from users and incorporating the results of data collection iteratively into product development. The result is a cyclical process of designing for usability then testing for usability, as shown in Figure 1. This process eventually satisfies the twin goals of first, meeting the needs of the product user and second, minimizing problems in product use.

Insert Figure 1 Here

Introduction to User-Centered Assessment Design

A UCAD approach is characterized by the following three features which are shared with UCD:

- Test design is primarily driven by an understanding of user capabilities and the context of score use;
- Test “performance” is evaluated using user satisfaction and user benefit in addition to traditional criteria of validity and reliability;
- Test design is interactive between developer and user.

In UCAD, we apply the concept and process of UCD to test design. At the center of UCAD is the user of test score information. Test design under UCAD seeks to answer a set of questions focusing on identifying this user and the context in which the test score information will be used.

- Who are the users of test score information?
- What are the users' tasks and goals?
- What information might users need, and in what forms do they need it?
- What prior knowledge do users have regarding test scores and their use?
- How can the design of test score information help users reach their goals?

Who is the user referred to in UCAD? Education includes different categories of users, including teachers, administrators, parents, and students. In other environments, users could be employers or mental health professionals. Whatever user category is the focus of assessment use, the defining characteristic of a user-centered approach is the placement of the user at the center of the methodology. For example, if the user is the teacher, test design would be driven primarily by understanding teacher capabilities and the context within which teachers will use test score information. Alternatively, if the user is the student, test design is likely to be different because student capabilities and the context within which students will use test score information are likely to be different.

In addition to the user, UCAD focuses on test score information. The focus is on test score information rather than on tests because the information derived from test scores is used—not the tests themselves (except, perhaps, to level a desk). Tests can take many forms; for example, computer-based or paper-based, multiple-choice questions, constructed-response questions or mixed-format, and timed or untimed. But UCAD is interested in these features only if they impact user benefit and user satisfaction.

Stages in User-Centered Assessment Design

Under UCAD, the design of assessments is user-centered and interactive. The interactive nature of test design is illustrated in the diagram of test development stages shown in Figure 2. The initial design of an assessment involves a feedback loop in which user needs are researched, prototype assessments are built and test score information is evaluated for usability. Following the feedback loop, assessment development follows a traditional pattern. This section walks through the stages in UCAD. Each stage is illustrated using the development of an assessment designed to provide formative information to teachers.

A test developer who has adopted UCAD would initiate test design by forming an initial understanding of the capabilities, needs, and goals of the focal user group and the context in which the test results would be used. For example, teachers whose only instructional strategy is to re-teach a unit use a different kind of assessment information than teachers who have a number of different instructional strategies available for use with students with different misconceptions or skill deficits. Similarly, the type and use of assessment information may also be influenced by the instructional context, such as the typical classroom size in which a teacher works, or the kind of learning targeted by the teacher. Or a particular type of assessment information might be better suited to inform instruction of a procedure (e.g., subtraction of multi-digit numbers), whereas another type of assessment information may be better suited to inform the teaching of conceptual understanding, such as the structure of the U.S. government.

Research on the capabilities of the focal user group and the context in which the test results would be used might use a number of different methodologies, including protocol analysis, interviews and focus groups. The subjects of the research are the intended users of the test score information, e.g., perhaps teachers rather

than test takers. If teachers are the intended users, teachers might be interviewed, after receiving a prototype assessment report in their classroom, regarding the usefulness of that report. Alternatively, a facilitator might lead a focus group of teachers as they explore scenarios of test score use.

After requirements have been gathered, an assessment prototype is developed and tested. The prototype being tested is not a set of items nor a score report but rather the score information that is derived from the assessment. The prototype is evaluated using teachers working within a classroom context to provide evidence that the design meets the goals. The centrality of test score information in UCAD is underscored by noting that a prototype could be developed by mocking up score reports that communicate test score information even in the absence of an actual test. But the focus of the evaluation is on the information provided in the report not in the report itself.

Assessment design may involve several cycles of design-evaluate-redesign iterations. The evaluation results are compared against the usability goals following each cycle. The purpose of iterating assessment design is to produce an assessment that offers information that is more effectively and efficiently used by teachers. Test development proceeds only after evaluation indicates that the assessment design sufficiently satisfies the usability goals.

After usability goals have been achieved, assessment development proceeds in a traditional fashion. Items or tasks are developed, reviewed, and field tested, and an operational assessment is then constructed and administered. Results are reported to parents, teachers and other educators. Reliability and validity evidence is collected and documented.

But UCAD continues after operational testing is complete and scores are reported. The reporting of assessment information provides an opportunity to collect additional usability evidence. Evidence is collected from teachers and other users on how effectively and efficiently test score information is used. Data may be collected through formal interviews and surveys in addition to the anecdotal reports that test developers frequently receive. This data is incorporated into assessment planning to improve the assessment design.

Pilot Study in Science Assessment

In this section, the design of a formative science assessment for grades 3 to 8 is used to illustrate UCAD. The test designers intended to develop an assessment that would provide teachers information that could be used to inform instructional decisions. This illustration describes only the initial stage of UCAD in which user needs were researched, prototype assessments were built and test score information was tested. The test designers collected data on science teachers and their classrooms to understand potential users of test score information and model likely contexts in which the test score information would be used. The data were collected in two stages: a broad survey of teacher needs and classroom practices and a more detailed interview to further explore issues raised by the survey.

Survey

In September of 2007, a survey was conducted with science educators from four states. The participants were classroom teachers, science curriculum coordinators, school administrators and state department of education science personnel. There were 61 respondents, 36 of whom were teachers and 25 of whom were science curriculum specialists or administrators. A more detailed description of the roles of the survey respondents is provided in Table 1. Furthermore, more than a quarter of the respondents worked in school

districts with more than 25,000 students. The percent of respondents who worked in school districts of different sizes is shown in Table 2.

Insert Tables 1 and 2 Here

The survey consisted of 18 items. The items asked about what grades should be assessed, what content should be included and what item formats to include. The survey items are shown in Appendix A.

Telephone Interviews

After reviewing the results of the survey, in-depth interviews were conducted in the winter of 2007 to help clarify some of the responses made on the survey. Teachers indicated on their survey form if they were willing to participate in a subsequent interview. Seven people who had completed the initial survey agreed to be contacted for a phone interview. Demographic information about all seven interview participants is provided in Table 3.

Insert Table 3 Here

The interview was designed to gather more detailed data with regard to survey questions for which additional, clarifying information was needed. There were five areas of interest in the follow-up interviews. Each area related to criteria for designing formative science assessments. Participants were also asked to describe what they liked and disliked about their current science assessments. The five areas of interest were:

- 1) Participants' thoughts on the use of grade bands in the design and delivery of formative assessments:

- What were their initial thoughts about the possibility of using grade bands?
 - Would participants prefer assessments on a grade-by-grade basis? If so, why?
- 2) Whether or not participants understood the meaning of the term “computer-based interactive item”
- 3) Whether or not the design of the formative assessment should include:
- A list of student misunderstandings
 - A proficiency scale
 - A forecast of how individual students would likely do on the summative, high-stakes state test
- 4) What type(s) of Professional Development were most efficacious:
- What should be the focus of Professional Development?
 - What types of Professional Development have participants had in the past that they have benefited from?
- 5) What types of science assessments do the participants currently use:
- What do people like and dislike about their current assessments?
 - Where did the items come from that are utilized in these assessments?

Results

The results are presented separately below for the survey and the telephone interview.

Survey

On the survey, teachers and administrators showed a similar pattern of results. Therefore, this section will discuss results aggregated across teachers and administrators. However, Appendix A presents results for each survey item disaggregated for teachers and administrators.

Results from the survey show that respondents want a science assessment that has the following features:

- Content that aligns at least 90% to the state content standards;
- A mix of item formats that reflects the mix of item formats on the NCLB state test;
- Item formats that include multiple-choice, short constructed-response and computer-based interactive items;
- Test development that meets the same technical standards as the NCLB state test;
- Information from test results that shows student misunderstandings, percent of the items scored correct, proficiency level, and profile of subtest scores;
- Professional development materials associated with assessment results.

The survey results provided conflicting findings on whether assessments should assess individual grades and courses or grade bands. More than three-quarters of respondents thought that individual grades and courses should be assessed by a formative assessment. However, more than half of respondents also thought that formative assessments should be based on grade bands. These apparently conflicting responses were explored further in the telephone interviews.

Telephone interviews

The results from the telephone interviews extend the findings presented for the survey. The questions asked in the telephone interviews were designed to clarify some of the responses made on the survey. Results from the interviews supported the survey findings that respondents wanted a science assessment with the following features:

- Information on the frequency and nature of student misunderstandings, perhaps organized around “big ideas” in science. Information on student proficiency level was not enough because it wouldn’t help the teacher know which concepts students who were not proficient needed help with.
- Professional development associated with the assessment results. Respondents indicated that they wanted professional development in the form of an extended-time, inquiry-based workshop in which teachers could interact with other educators and with the materials and lessons they will be using with students.
- A reliable forecast of how students are likely to perform on the NCLB test.
- Rapid scoring and reporting of individual student results and classroom results. Ideally, reports would be provided within 24-hours of testing and would offer a range of analyses including item analysis, class item analysis and performance aggregated by IEP status.

As noted earlier, survey, findings on whether assessments should assess individual grades and courses or grade bands were conflicting. During the interviews, questions were asked that attempted to resolve this conflict. Most respondents in the interviews said that their state’s science curriculum was tightly defined to include certain topics and exclude others on a grade-by-grade basis so that grade bands would be counter-productive.

Researchers were unsure what survey respondents wanted when they responded that they favored a science assessment that included computer-based interactive items. During the interviews, most respondents correctly knew that a computer-based interactive item meant the student could interact with the computer to document change (e.g., quantity, physical state, direction, speed, mass, displacement).

Prototype

The last step in the initial stage of UCAD in which user needs are researched, prototype assessments are built and test score information is tested, is to build and test a prototype of test score information that is intended to meet the needs of users as identified via the results of the earlier survey and phone interviews. At this point in the project, we would develop a prototype score report specifically for teachers. This prototype would then be presented to teachers who would evaluate the usefulness of the information reported.

Conclusion

To conclude this paper, we compare UCAD to conventional test development and evidence-centered design (ECD). We argue that UCAD, in conjunction with what is best in conventional test development and ECD, can improve the assessments that are developed.

Conventional test development

Test score use is an aspect of test design under conventional test development. But conventional treatments of test score use focus on test purpose and domain of inference rather than on the abilities of test score users and the context of test score use. According to Millman and Greene (1989), test purpose corresponds to the category of educational decisions in which test scores are anticipated to play a role. They identify three domains in which test scores are used to help make decisions: the curricular domain, the cognitive domain and a future criterion setting (Millman & Greene, 1989). Nearly 20 years later, Schmeiser and Welch (2006) discuss test score use with reference to the domain to which test score inferences are to be made. Their view emphasizes that accurate construct interpretation is prerequisite to defensible test score use.

The *Standards for Educational and Psychological Testing* (1999) state that the test developer is responsible for providing rationales and evidence supporting advocated test use. However, the evidence and rationales supporting test score use cited in the *Standards* generally center on the need to establish that construct interpretation that supports test score use. For example, Standard 1.1 in the validity discussion states: “A rationale should be presented for each recommended interpretation and use of test scores, together with a comprehensive summary of the evidence and theory bearing on the intended use or interpretation” (p. 17). Discussions of the kind of rationale, evidence and theory that is relevant to intended test score use reference inferences to constructs and domains. Similarly, in the discussion of test development, the *Standards* defend test purpose by reference to the test domain. For example, Standard 3.2 states: “The purpose(s) of the test, definition of the domain, and the test specifications should be stated clearly so that judgments can be made about the appropriateness of the defined domain for the stated purposes of the test ...” (p. 43).

Conventional test development may be characterized as more test-centered than UCAD. A test-centered approach toward test-development is characterized by three features:

- Innovation is driven by technical developments;
- Test “performance” is evaluated using measurement characteristics;
- Test development tends to be sequential.

Evidence-centered design

Under ECD, test development is viewed as an evidentiary argument for reasoning from observed student performance in a few particular circumstances to what students know, can do, or have accomplished in a broader context (Mislevy, Steinberg, & Almond, 2003). The design framework is based on the principles of evidentiary reasoning. The framework is intended to ensure that the way in which evidence is gathered and

interpreted bears on the underlying knowledge and purposes the assessment is intended to address. The stages and structures in test development are designed to build on a sound chain of reasoning from what we observe to what we infer. Such a framework not only makes the underlying evidentiary structure of an assessment more explicit, but it makes the operational elements of an assessment easier to reuse and to share.

ECD provides a conceptual design framework intended to make item and task development more efficient by organizing test development around the argument for making inferences from student performance to statements of achievement (Mislevy, Steinberg, Almond & Lukas, 2006). Several features of ECD recommend the approach as the foundation for improved test development. First, ECD makes item and task development more efficient by incorporating generative, reusable, and duplicable features in the tools used for test development. Second, ECD is a way to systematically and scientifically develop assessments to give evidence of construct validity. ECD ensures evidence for construct validity through the test development process. Under ECD, the way assessment evidence is gathered (test design) and interpreted (scoring, calibration, and scaling) bears on the underlying knowledge and skills the assessment is intended to address, i.e., construct validity.

The core of ECD is the five basic models that comprise test development (Mislevy & Riconscente, 2006):

- **Student Model.** A student model describes characteristics of students, typically latent variables, that we wish to model. These characteristics are the student knowledge, skills and abilities (KSA's) that are the targets of inference.
- **Evidence Model.** The evidence model describes how to extract evidence of latent variable status from student performance. The evidence model makes explicit the relationship of student

performance (observable variables) to the underlying construct (latent variables in the student model).

- **Task Model.** A task is best presented as an activity that is broadly defined to include context such as instructions, passages and response method. A task model specifies characteristics of a task that should be included in the evidence model. Task model variables influence the latent variables that are in the student model.
- **Assembly Model.** The assembly model describes the mixture and sequence of tasks. The assembly model would appeal to the information in the task model to select and organize the sequence of tasks. A CAT algorithm is an example of an assembly model; so are test specifications for producing a series of content- and statistically parallel fixed-form tests.
- **Delivery Model.** The delivery model describes the interface between the student and the assessment. From an operational perspective, the delivery model specifies requirements and data structures for the way students will interact with tasks, including how materials are provided, tools and affordances operate, and work products are captured.

Within ECD, the construct holds center stage—but user satisfaction and user benefit are not explicitly addressed. For example, the five basic models that constitute the core of ECD do not include a user model. ECD may be characterized as more construct-centered. Test development under a construct-centered approach is organized by a strong program of validity. A strong program of validity was outlined by Loevenger (1957) and Messick (1989), among others. Under a strong program of validity, construct theory dominates every aspect of test development. Answers to questions of item content, scoring model, and validity research are largely deduced from construct theory.

A Conjunctive Approach

Neither conventional test development nor ECD should be understood as deficient compared to UCAD, only different in emphasis. The differences in emphasis among conventional test development, ECD, and UCAD can become a strength if what is best about each approach can be combined in a conjunctive approach that results in improved assessment score interpretation and use. What UCAD brings to a conjunctive approach is to incorporate into the test development process the perspective of those most directly at the forefront of the educational endeavor: teachers, students, administrators, parents. The inclusion of these perspectives throughout the development process can help to build a gestalt that combines the strengths tapped so well by the classical and ECD models of test development with those of the experts in the local, school and state level academic needs and requirements.

The most straightforward path toward such a combined approach would seem to be an additive one in which, as illustrated earlier in Figure 2, the initial design-evaluate-redesign iterative process is incorporated as one of the earliest steps in the test development process, ideally in concert with development/finalization of item and test specifications. This is also, of course, one of the central tenets of a user-centered design approach; early, direct input from—and incorporation of—primary user concerns into the test design specifications provides an ideal mechanism that will:

- focus the spotlight of the test on issues of greatest concern at the teaching-learning nexus
- foster a strong sense of ownership and utilitarianism among teachers
- further customize formative assessments and the diagnostic information resulting from them to meet local needs

In addition to the initial incorporation of the design-evaluate iterative process at the beginning of test development, a conjunctive approach that includes a focus on user input would also incorporate user

perspectives as possible throughout the rest of the development process. Note that, to some extent, good test development process already often incorporates the perspectives of local educators and other experts through such mechanisms as item and test form content reviews, sensitivity reviews, standards setting/review meetings, and similar forums. However such events, though valuable and contributory to an inclusive approach, are usually directed at more specific elements of the development process and rarely provide opportunities for these users to impact directly the shape, content, or output of the tests themselves.

A more formal approach to continuing the inclusiveness of the initial incorporation of user input during the design-evaluate iteration phase of test development is to incorporate a mandatory step near the back end of the test development process as shown in Figure 2, that solicits feedback from relevant users of such primary materials as the test forms, answer documents, score reports, and other ancillary documents, as well as surveys to collect users' overall impressions of each set of materials and the products and process overall. In effect, this additional formal opportunity for review constitutes yet another cycle of evaluation in which users now have the chance to determine the extent to which the test developer has been able to incorporate the users' perspectives into the testing program as a whole, and to express any new or lingering concerns or suggestions in preparation for the next generation of the program.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Bevan, N., & Schoeffel, R. (2001). A proposed standard for consumer product usability. In The Proceedings of the 1st International Conference on UAHCI, New Orleans, August, 2001.
- Brown, B. A. T., Geelhoed, E., & Sellen, A. J. (2001). The use of conventional and new music media: Implications for future technologies. In M Hirose (Ed.), *Human-Computer Interaction*. IOS Press: Amsterdam, The Netherlands.
- Hackos, Joann T. and Redish, Janice. *User Interface Task Analysis*. John Wiley & Sons. 1998
- Lamberts, H. (2002). Case Study: A PDA Example of User Centered Design. *Lecture Notes In Computer Science*; Vol. 2411. Proceedings of the 4th International Symposium on Mobile Human-Computer Interaction. 329-333
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635-694.
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). Washington, DC: American Council on Education and National Council on Measurement in Education.
- Mislevy, R. J., Steinberg, L. S., Almond, R. G., & Lukas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, I. I. Bejar & R. J. Mislevy (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 15-47). Mahwah, NJ: Erlbaum.
- Mislevy, R. J., & Riconscente, M. M. (2006). Evidence-centered assessment design: Layers, concepts, and terminology. In S. Downing & T. Haladyna (Eds.), *Handbook of Test Development* (pp. 61-90). Mahwah, NJ: Erlbaum.

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). On the structure of educational assessments.

Measurement: Interdisciplinary Research and Perspectives, 1, 3-67.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman & S. W. Draper (eds.), User-centered system design: New perspectives on human-computer interaction (pp. 32 – 65). Hillsdale, NJ: Erlbaum.

Norman, D. A. (1988) *The Psychology of Everyday Things*. Basic Books.

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Size of School District	Less than 1,000	Between 1,000 and 2,499	Between 2,500 and 4,999	Between 5,000 and 9,999	Between 10,000 and 24,999	Greater than 25,000	Totals
All	8%	15%	9%	23%	17%	28%	53
Teachers	11%	17%	6%	23%	14%	29%	35
Curr/Admin	0%	11%	17%	22%	22%	28%	18

Table 1. Frequency of responses to the survey question: "What is the size of your school district?"

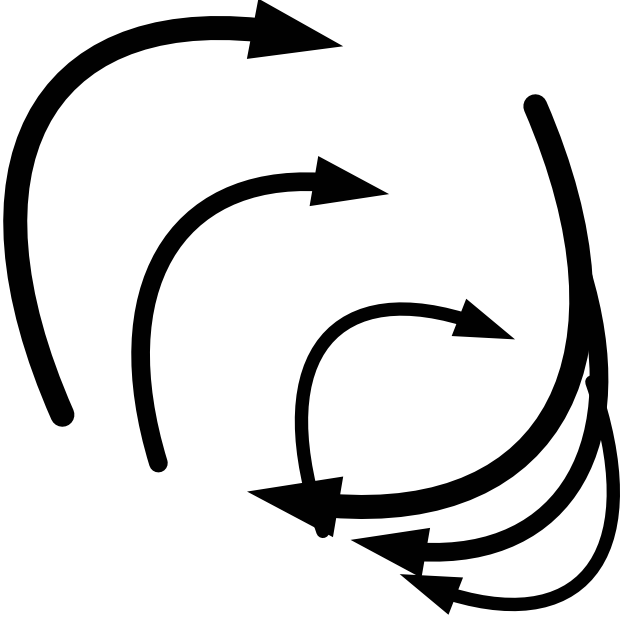
Title	Number	Teacher	Curr/Admin
Building-level administrator	5		5
Content coordinator	1		1
Curriculum specialist	1		1
District-level administrator	1		1
Elementary teacher	8	8	
High school science teacher	15	15	
HS science curriculum coordinator	1		1
Middle school curriculum coach	1		1
Middle school science teacher	11	11	
Other teacher	2	2	
Science coach	1		1
Science coordinator	2		2
Science implementation specialist	3		3
Science mentor	1		1
Science specialist	7		7
SDE	1		1
Totals	61	36	25

Table 2. Frequency of responses to the survey question: “What is your role in your school district?”

Participant	Gender	State	Duties	Additional Comments
1	F	NC	Teaches 9 th , 10 th , 11 th and 12 th grade science	Teaching for 9 years. Her school has roughly 1,300 kids enrolled
2	F	MN	Teaches K – 6	Her school has a split curriculum, her duties are related to science mostly
3	M	NC	Former science teacher, now High School assistant principal	
4	F	SC	K - 12 Science and Social Studies Coordinator	National Board Certified Teacher in Early Adolescent Science and Masters in Early Childhood Development
5	F	NC	Middle School Administer	Taught high school science. Was a science curriculum specialist for two years
6	F	SC	Science coach in SC elementary school	National Board Certified Teacher in General Studies
7	F	MN	Teaches 7 th – 9 th science	Teaches two sections of life science, two of earth science, and one of physical science

Table 3. The participants in the telephone interviews.

Figure 1.



Appendix A

Survey Items and Results

1) Should individual grades/courses be assessed by a formative science assessment?

Should individual grades/courses be assessed by a formative science assessment?	Yes	No	Totals
All	78%	22%	63
Teachers	72%	28%	36
Curr/Admin	88%	12%	25

2) If you selected "yes" on question 1, please rank the grades/courses from most important (#1) to least important (#10). If you selected "no" on question 1, please skip to question 3.

3) Should there be a common set of formative science assessments based on grade bands? For example, grades 3-5 and grades 6-8 would represent possible grade bands.

	Yes	No	Totals
Grade Bands--All	59%	41%	63
Grade Bands--Teachers	53%	47%	36
Grade Bands--Curr/Admin	68%	32%	25

4) If you selected "yes" on question 3, please rank the grade bands that should be assessed from most important (#1) to least important (#8). If you selected "no" on question 3, please skip to question 5.

5) How much does the content of the formative science assessment need to align to your state test?

Content Aligned to State Test	1-100%	2-At least 90%	3-At least 80%	4-At least 70%	5-At least 60%	6-At least 50%	7-Less than 50%	Totals
All	68%	19%	11%	2%	0%	0%	0%	63
Teachers	56%	25%	17%	3%	0%	0%	0%	36
Curr/Admin	88%	8%	4%	0%	0%	0%	0%	25

6) How important is it for a formative assessment in science to test content taught at prior grade levels?

Test Prior Content	1-Very important	2-Moderately important	3-Somewhat important	4-Not important	Totals
All	33%	31%	26%	10%	61
Teachers	35%	29%	29%	6%	34
Curr/Admin	32%	32%	24%	12%	25

7) What item types are most desirable on a formative science assessment? Rank as 1 the most desirable item type and rank as 5 the least desirable item type.

All	1-Multiple choice	2-Short constructed response	3-Extended constructed response	4-Passage-based questions	5-Computer-based interactive items	Totals
Rank 1	39%	26%	5%	11%	19%	62
Rank 2	13%	23%	11%	24%	29%	62
Rank 3	13%	20%	18%	23%	26%	61
Rank 4	9%	29%	28%	26%	9%	58
Rank 5	29%	4%	41%	14%	13%	56

Teachers	1-Multiple choice	2-Short constructed response	3-Extended constructed response	4-Passage-based questions	5-Computer-based interactive items	Totals
Rank 1	43%	23%	6%	14%	14%	35
Rank 2	14%	23%	6%	23%	34%	35
Rank 3	11%	20%	20%	23%	26%	35
Rank 4	9%	31%	23%	29%	9%	35
Rank 5	24%	3%	48%	9%	15%	33

Curr/Admin	1-Multiple choice	2-Short constructed response	3-Extended constructed response	4-Passage-based questions	5-Computer-based interactive items	Totals
Rank 1	32%	28%	4%	8%	28%	25
Rank 2	12%	24%	16%	24%	24%	25
Rank 3	17%	21%	17%	25%	21%	24
Rank 4	10%	24%	38%	19%	10%	21
Rank 5	33%	5%	29%	24%	10%	21

8) How important is it that the mix of item types on a formative science assessment matches the mix of item types on your state test?

How important is it that the mix of item types on a formative science assessment matches the mix of item types on your state test?	1-Very important	2-Moderately important	3-Somewhat important	4-Not important	Totals
All	63%	23%	11%	3%	62
Teachers	56%	28%	14%	3%	36
Curr/Admin	75%	13%	8%	4%	24

9) On a formative science assessment, how important is it that the formative assessment be developed with the same standards of expert review, field testing, reliability and validity as your large-scale state test?

On a formative science assessment, how important is it that the formative assessment be developed with the same standards of expert review, field testing, reliability and validity as your large-scale state test?	1-Very important	2-Moderately important	3-Somewhat important	4-Not important	Totals
	All	86%	10%	5%	0%
Teachers	89%	6%	6%	0%	36
Curr/Admin	80%	16%	4%	0%	25

10) What kind of scores would you like to receive from a formative science assessment? Please rank the kind of scores you would like to receive. Rank as 1 the most important kind of score and rank as 7 the least important kind of score.

Kinds of Scores--All	1- Percent correct	2- Percentile rank	3- Proficiency level	4-Forecast of achieving proficiency on the end-of-year test	5- Student strategies	6-Student misunderstandings	7- Profile of subtest scores	Totals
Rank 1	29%	5%	14%	10%	5%	25%	13%	63
Rank 2	13%	3%	19%	14%	16%	19%	16%	63
Rank 3	8%	8%	16%	10%	17%	24%	17%	63
Rank 4	6%	11%	24%	11%	21%	15%	11%	62
Rank 5	18%	15%	12%	23%	13%	7%	12%	60
Rank 6	17%	26%	9%	14%	17%	9%	9%	58
Rank 7	9%	32%	4%	19%	11%	4%	23%	57

Kinds of Scores--Teachers	1- Percent correct	2- Percentile rank	3- Proficiency level	4-Forecast of achieving proficiency on the end-of-year test	5- Student strategies	6-Student misunderstandings	7- Profile of subtest scores	Totals
Rank 1	31%	6%	8%	8%	6%	28%	14%	36
Rank 2	11%	6%	22%	11%	14%	25%	11%	36
Rank 3	11%	8%	19%	17%	14%	17%	14%	36
Rank 4	6%	14%	33%	11%	19%	11%	6%	36
Rank 5	14%	26%	9%	17%	14%	6%	14%	35
Rank 6	15%	18%	9%	18%	21%	9%	12%	34
Rank 7	12%	24%	0%	21%	12%	6%	26%	34

Kinds of Scores-- Curr/Admin	Percent correct	Percentile rank	Proficiency level	Forecast of achieving proficiency on the end-of-year test	Student strategies	Student misunderstandings	Profile of subtest scores	Totals
Rank 1	24%	4%	24%	12%	4%	20%	12%	25
Rank 2	16%	0%	12%	20%	16%	12%	24%	25
Rank 3	4%	8%	12%	0%	24%	36%	16%	25
Rank 4	8%	8%	13%	8%	25%	17%	21%	24
Rank 5	26%	0%	13%	35%	9%	9%	9%	23
Rank 6	18%	41%	9%	5%	14%	9%	5%	22
Rank 7	5%	38%	10%	19%	10%	0%	19%	21

11) On a formative science assessment, how important is it that professional development materials accompany the assessment?

On a formative science assessment, how important is it that professional development materials accompany the assessment?	1-Very important	2-Moderately important	3-Somewhat important	4-Not important	Totals
All	87%	11%	2%	0%	63
Teachers	89%	11%	0%	0%	36
Curr/Admin	84%	12%	4%	0%	25

12) What features are most important on a formative science assessment? Please rank the features that you would like to see on a formative science assessment ranking as number 1 the most important feature and as number 5 the least important feature.

Ranking of Features-- All	The content aligns 100% to state test	Tests content taught at prior grade levels	The mix of item types matches the mix on state test	The assessment has same standards as state test	Prof. Dev. materials accompany assessment	Totals
Rank 1	60%	3%	6%	27%	3%	62
Rank 2	11%	8%	45%	18%	18%	62
Rank 3	13%	15%	20%	30%	22%	60
Rank 4	8%	23%	15%	17%	37%	60
Rank 5	8%	50%	13%	8%	20%	60

Ranking of Features-- Teachers	The content aligns 100% to state test	Tests content taught at	The mix of item types matches	The assessment has same	Prof. Dev. materials accompany	Totals
-----------------------------------	---------------------------------------	-------------------------	-------------------------------	-------------------------	--------------------------------	--------

		prior grade levels	the mix on state test	standards as state test	assessment	
Rank 1	51%	6%	9%	31%	3%	35
Rank 2	14%	14%	43%	20%	9%	35
Rank 3	15%	9%	26%	26%	24%	34
Rank 4	9%	29%	11%	14%	37%	35
Rank 5	11%	43%	11%	9%	26%	35

Ranking of Features— Curr/Admin	The content aligns 100% to state test	Tests content taught at prior grade levels	The mix of item types matches the mix on state test	The assessment has same standards as state test	Prof. Dev. materials accompany assessment	Totals
Rank 1	72%	0%	4%	20%	4%	25
Rank 2	8%	0%	48%	16%	28%	25
Rank 3	13%	25%	8%	33%	21%	24
Rank 4	4%	17%	22%	22%	35%	23
Rank 5	4%	57%	17%	9%	13%	23

13) How many different science assessments do you use to monitor your students' achievement in science (e.g., the state science test and a particular formative science test would be "2")?

Number of Assessments	Frequency
0	15
1	13
2	10
3	9
4	7
5	6
25	2
7	1
8	1
12	1

14) Please list the different science assessments (up to 5) from question 13 by name below:

Frequency	Name or Description of Science Assessment
37	State Test
19	Benchmark Test
10	Teacher Created Assessments
8	Grade Level Test
7	Project
6	MAP (NWEA Assessment)
5	Textbook Test

4	Lab Tests
4	Topic Assessment
3	End of Course Test
3	Performance Test
2	Class Exams
2	Observation
2	Pre and Post Tests
2	Quizzes
1	Ask Project Formative Science Tests
1	Classroom Experiments
1	Computer Based Assessment
1	Computer Test
1	Constructed Response
1	County Mid Term Exams
1	Cumulative Tests Every 3 Weeks
1	Curriculum Tests
1	District Quarterlies
1	Essential Questioning
1	Informal Assessments
1	Midterm
1	Multiple Choice
1	National Science Test
1	Notebook Entries of Students
1	Questioning And Discourse in the Classroom
1	Sample MCA Test
1	School Common Exams
1	Science Journals

15) What is the size of your school district?

Size of School District	1-Less than 1,000	2- Between 1,000 and 2,499	3- Between 2,500 and 4,999	4- Between 5,000 and 9,999	5- Between 10,000 and 24,999	6- Greater than 25,000	Totals
All	8%	15%	9%	23%	17%	28%	53
Teachers	11%	17%	6%	23%	14%	29%	35
Curr/Admin	0%	11%	17%	22%	22%	28%	18

16) What is your role in your school district?

Title	Number	Teacher	Curr/Admin
Building-level administrator	5		5
Content coordinator	1		1
Curriculum specialist	1		1
District-level administrator	1		1
Elementary teacher	8	8	
High school science teacher	15	15	
HS science curriculum coordinator	1		1
Middle school curriculum coach	1		1
Middle school science teacher	11	11	
Other teacher	2	2	
Science coach	1		1
Science coordinator	2		2
Science implementation specialist	3		3
Science mentor	1		1
Science specialist	7		7
SDE	1		1
Totals	61	36	25

17) As part of our ongoing research effort, we may contact some survey participants for follow-up telephone interviews (who will be offered a stipend for their time). Are you willing to participate in a 45 to 60 minute phone interview to answer a few more questions about formative science assessments?

18) If you answered "yes" to question 17, please give us the following information:

Appendix B
Interview Questions

Question 1: For science, we can create the assessments so that they can be delivered in “grade bands.” For example, we could create a grade band for grades 3 – 5. This means that 3rd, 4th and 5th graders would all be getting 3rd, 4th and 5th grade science items. Grade bands allow for flexibility in scope and sequence when teaching the curriculum.

- a. What are your initial thoughts on using grade bands?
- b. Would you prefer to have assessments per grade? If so, why?

Question 2: When you hear the term “computer-based interactive items,” what does it mean to you?

Question 3: There are several things we can offer you after students take the formative science assessment that should give you a better understanding of the students’ science performance.

- a. For example, after a student takes the Science assessment, we can provide you with a list of student “misunderstandings” for the areas that the student incorrectly answered on the test. What would you expect to see for student misunderstandings?
- b. We can offer you data related to a student’s proficiency level. What do you currently use to measure proficiency? What does proficiency mean to you?
- c. How important is it for you to get a forecast of how a student will perform on the high-stakes tests after they take a formative test?

Question 4: What types of Professional Development would you and/or your teachers like to participate in and benefit the most from (e.g., workshops, conferences, lesson-study teams, Professional Learning Teams, Learning Coaches, online courses)?

- a. What should be the focus of the Professional Development?
- b. What kind of Professional Development have you previously had that was beneficial to you and/or your science teachers?

Question 5: What science assessments do you currently use?

- a. What do you like and dislike about those assessments?
- b. Where did the items come from (e.g., locally developed items, textbook items, vendor developed)?

Additional Comments:

Participants were asked for anything additional they would like to add.