

## **BALANCING VARIED ASSESSMENT FUNCTIONS TO ATTAIN SYSTEMIC VALIDITY: THREE IS THE MAGIC NUMBER**

**Daniel T. Hickey\*, Steven J. Zuiker\*, Gita Taasobshirazi\*\*,  
Nancy Jo Schafer\*\*\*, and Marina A. Michael \*\*\*\***

\*Learning Sciences Program, Indiana University, USA

\*\*Department of Educational Psychology and Instructional Technology, University of Georgia, USA

\*\*\*Department of Early Childhood Education, Georgia State University, USA

\*\*\*\*Combined Program in Cognitive, Developmental and Educational Psychology,  
University of Cyprus, Cyprus

### Abstract

Accountability-oriented reforms demand immediate *and* continual gains on achievement test, for all students, and without diminishing other outcomes or undermining instruction. This paper describes a framework for aligning classroom assessment and external testing with the aim of negotiating these seemingly contradictory goals. The framework varies the sensitivity to instruction and the representations of knowledge across approaches to assessment. Cycles of design-based studies successively refine relationships between a curriculum and the frame that each assessment provides. Doing so, we argue, leverages the unique formative and summative balance across assessments in order to scaffold learning and demonstrate the "consequential" validity of our strategy without compromising curricula, instruction, or the "evidential" validity that warrants their continued use.

Achievement tests present an enduring challenge for education. Testing-related tensions have been exacerbated by accountability-oriented educational reforms that mete out rewards and sanctions to schools based on test scores. The *No Child Left Behind* (NCLB) act in the United States demands annual gains on criterion-referenced achievement

tests for most subgroups of students in every school. Meanwhile, the broader community expects that these gains not come at the expense of other desirable educational outcomes (e.g., scores on other non-targeted tests, graduation rates, future educational success, etc). Many consider these goals contradictory.

While we certainly have strong opinions about the intentions and conduct of test-based accountability, we take an agnostic stand towards these practices. We instead direct our concern towards the many ineffective and counterproductive responses to them (Popham, 2003; Sloane & Kelly, 2003). In particular, we are concerned about "test-prep" programs that prepare students to pass specific tests. Many of these programs are computer-based and train students to recognize numerous specific associations. Because a few of the learned associations are sure to be helpful in recognizing correct and incorrect responses on the targeted test, the programs are being aggressively and successfully marketed to schools as "evidence-based" solutions to student achievement (e.g., Sleek, 2004; StudyIsland, 2004; Tingley, Thrall, & Ward, 2001; Ysseldyke & Tardrew, 2002).<sup>1</sup> Our review of the scant research literature on such practices convinced us that these yield trivial gains on the targeted tests, with little or no evidence of generalized student learning.<sup>2</sup> Of particular concern is evidence that widely-touted success at directly increasing criterion-referenced scores is associated with decreased scores on norm-referenced tests covering the same content (e.g., Hoff, 2004; Markley, 2004; Schemo & Fessenden, 2003).

Implicit in test-based reform is "systemic validity" (Frederiksen & Collins, 1989). Systemic validity is achieved when desired educational outcomes are attained by virtue of assessing those same outcomes. We contend that most responses to test-based accountability are not systemically valid. This article introduces an alternative response that aligns formative classroom assessments with external achievement tests. This model of assessment practice emerged in two consecutive studies involving the *GenScope* introductory genetics software (Horwitz & Christie, 2000). The *GenScope* software as well as the studies described here were all funded by the U. S. National Research Foundation. The approach was further refined in a recently-completed study of NASA-sponsored science curriculum, and is being applied in two recently funded studies (one in middle school ecology and one in elementary mathematics).

Thanks in part to comprehensive reviews by Black and Wiliam (1998), and the US National Research Council (2001a), the potential of formative classroom assessment is now widely recognized. These reviews show that prior research on formative assessment has focused on *indirectly* improving student knowledge, via feedback to teachers (c.f., Anderson, 2003). Our new approach builds on this work, but also extends it by using classroom assessments to more directly impact student learning. In this regard, our approach draws strongly on Duschl and Gitomer (1997), whose portfolio-based "assessment conversations" highlighted the potential of what we call a "discursive" approach to formative feedback.

We also build on and extend research aiming to "bridge the gap" between classroom assessment and external testing (e.g., NRC, 2003; Wilson, 2004). We argue that the alignment of assessment and testing works best when three or more such "levels" of assessment are clearly defined. This conclusion is the basis of this article's title and the first "feature" of what we have come to call our "3 x 3" framework. The first part of the article outlines these multiple levels of assessment, and shows how each level affords very

different formative and summative potential. The second feature of our approach is the use of increasingly formal representations of domain knowledge across these levels. At increasingly distal levels, the representations of knowledge shift from "cultural" *discourse* to "cognitive" *understanding* to "behavioral" *achievement*. The second section of this paper summarizes these three ways of conceptualizing knowledge, and shows how embedding them in a "comparative" approach (Greeno, Collins, & Resnick, 1996) improves systemic validity within and across levels.

The third feature of our approach is the use of contemporary "design-based" research methods. Eschewing the traditional distinction between "basic" laboratory research and "applied" classroom research, design-based methods call for the development of "intermediate-level" theories across iterative refinements of practice. We argue that three increasingly formal cycles are needed to maximize systemic validity and attain meaningful gains in student achievement. As described in the third section of this article, we label these cycles *implementation*, *experimentation*, and *evaluation*. Finally, the fourth section summarizes the evidence obtained so far regarding the impact of this framework on student learning, and describes the two new studies getting under way.

### Increasingly Distal Levels of Assessment

The first feature of our approach concerns the various *levels* of assessment. The need to consider the relationship between different levels has been central to recent considerations of assessment. For example, the expert panel on student testing assembled by the National Research Council concluded:

...aspects of learning that are assessed and emphasized in the classroom should ideally be consistent with (though not necessarily the same as) the aspects of learning targeted by large-scale assessments. In reality, however, these two forms of assessment are often out of alignment. The result can be conflict and frustration for both teachers and learners. *Thus there is a need for better alignment among assessments used for different purposes and in different contexts...*" (NRC, 2001b, p. 3)

Our own prior efforts to align the formative value of "internal" classroom assessment to improve performance on "external" assessment revealed the limits of alignment involving just two levels (Hickey, et al., 2000). Our efforts to fine-tune the formative value of classroom assessments compromised the evidential validity of our external test. While this led to dramatically increased gains on the external test, it also afforded an unfair advantage to our implementation classes over the comparison classes (as reported in Hickey, Kindfield, et al., 2003). In conventional measurement terms, our efforts to maximize the "consequential" validity of our classroom assessments compromised the "evidential" validity of our external test. We concluded that when fine-tuning formative feedback at one assessment level in order to maximize performance at a second level, a third more-distal level of assessment is needed if one wishes to have a valid estimate of learning in a comparison curriculum targeting the same content standards.

As summarized in Table 1, we distinguished between five levels of educational outcomes, each signifying a greater "distance" from a particular curricular activity. These levels are labeled *immediate*, *close*, *proximal*, *distal*, and *remote*. The labels and specifications were taken directly from a summative analysis reported by Ruiz-Primo, Shavelson, Hamilton and Klien (2002; see Kennedy, 1999, for a similar analysis). These levels obviously define a continuum. We argue identifying discrete points on this continuum provide clarity that is essential for attaining systemic validity. For example, assessments at the distal and remote levels (e.g., criterion-referenced and norm-referenced achievement tests) are necessarily too removed from a specific curricular routine to meaningfully assess its value. But this same distance affords a preclusion of bias towards a particular curriculum, making such measures valid for some kinds of large-scale comparisons of entire curricula.

Ruiz-Primo, et al., (2002) defined their five levels in an influential external evaluation of massive efforts to reform science curriculum in urban US schools. At each of the five levels, they developed rigorous summative assessments of learning in the reform and comparison classrooms. While they found evidence of reform impact at the immediate, close, and proximal levels, the study provided convincing evidence that the reforms had no impact at the distal or remote levels. As outlined below, assessments at the immediate, close, and proximal levels are not independent of a given curriculum. As such, those assessments do not yield valid comparisons of learning when used with other non-targeted curricula. The lack of impact on "external" measures helped scuttle the reforms, and gave us an opportunity make a key point about our approach. We are now confident that the reforms studied by Ruiz-Primo et al. would have yielded achievement gains at the distal *and* remote levels if they had instead aimed to maximize the consequential validity (i.e., the formative value) of assessments at immediate, close, and proximal levels, as we show below.

One of our core arguments is that the dichotomy between "internal" "formative" assessments and "external" "summative" tests obscures the most powerful methods of attaining systemic validity. Rather, we assume that every assessment and every test has both formative and summative *functions*, but those functions are very different across different assessment levels. Thinking about the diverse (and potential) functions of every assessment and test provides many useful insights. Black and Wiliam (1998) review evidence showing that external tests undermine the formative potential of typical classroom assessments (see also NRC, 2003; Shepard, 2000). While this is certainly true, we also worry that the formative potential of a given assessment is often undermined by the more salient summative functions of that same assessment. Consider, for example, the most typical formative classroom assessment goals of enhancing teacher decision-making. From the perspective of the individual student, there is nothing particularly formative about classroom assessments that are used for curricular refinement, promotion, and classroom accountability. When learners do not recognize the formative potential of assessments, their value for directly impacting student learning is quite limited (Gipps, 1999).

Table 1: Five Levels of Assessment and their Ideal Formative Functions

Level	Primary Orientation	Relevant Time Scale	Ideal Assessment in our Framework	Principal Formative Function for Student	Principal Formative Functions for Others	Ideal Domain Knowledge Representation
Immediate	Enactment of Specific Curricular Routine	Minutes	Event-oriented observations (informal from the enactment of the curriculum)	A classroom's enactment of specific curricular routines and participation in domain knowledge practices.	A classroom's enactment of specific curricular routines and participation in domain knowledge practices.	Enacted knowledge practices, discourse during a specific curricular routine
Close	Specific Curricular Routines	Days	Activity-oriented quizzes (semi-formal classroom assessments)	Understanding of domain knowledge targeted by specific curricular routines.	Refining specific curricular routines and providing informal remediation to students.	Semi-formal representation, discourse around representations similar to the curricular routines.
Proximal	Entire Curricula	Weeks	Curriculum-oriented exams (formal classroom assessments)	Understanding of the domain knowledge targeted by an entire curriculum.	Formally refining entire curricula and providing formal remediation to students.	Formal representation of the concepts covered in the curriculum.
Distal	Regional or National Content Standards	Months	Criterion-referenced tests (external tests aligned to content standards)		Selecting curricula that have the largest impact on achievement in broad content domains.	Formal representation of associations drawn from the standards
Remote	National Achievement	Years	Norm-referenced tests (external tests standardized across years (ITBS, NAEP, TIMMS		Refining long-term impact of policies on broad achievement targets.	Formal representation of associations drawn from national samples of achievement.

In our efforts to define and balance functions within and across assessment levels, we have found it useful to define the primary *orientation* of each level, as shown in the second column of Table 1. For example, this reveals that the immediate-level is oriented towards *events* (e.g., the enactment of specific activities in a specific context).<sup>3</sup> This differs in important ways from the close level, which is oriented towards specific curricular *activities* (i.e., independent of the actual enactment of the activity), and from the proximal level, which is oriented towards a collection of activities in a particular curricula. Likewise, while the distal level is oriented towards standards (such as the cut scores on criterion-referenced tests, which change over time), the remote level is oriented towards changes in achievement over time.

In our efforts to broaden our understanding of the formative functions across different assessment levels, we have found Lemke's (2000) notion of *timescales* theoretically powerful. The timescale associated with each level is shown in the third column of Table 1. For example, close-level assessment is oriented towards specific learning activities. In order to assess learning from a specific lesson closely, it would likely be most appropriate to assess that learning through end of lesson sequence cycles, rather than an entire curriculum. As such, the relevant timescale for close-level assessment is the span of roughly days.

This notion of timescales is crucial to understanding and balancing the formative functions at different levels. Proximal levels resonate with shorter timescales that are necessary for directly advancing student learning, while more distal levels represent the longer timescales that are necessary for predicting longer term mastery and evaluating more comprehensive curriculum. As shown in the fourth column of Table 1, we use this orientation as part of the descriptive labels for what we believe are the ideal assessments at each level of our approach: *event-oriented observations*, *activity-oriented quizzes*, *curriculum-oriented exams*, and *criterion-referenced tests*, and *norm-referenced tests*.

The consequences of the preceding insights in our work are captured in the fifth and sixth columns of Table 1. These columns show that the formative functions of assessment are different at each level, and that the formative functions for students within each level are different than those for teachers (and/or administrators, researchers, and policy makers). Ours is certainly not the first consideration of the varied formative functions of assessment. The NRC reports on classroom assessment (2001a), student testing (2001b) and the alignment of classroom and external assessment (2003) reviewed numerous studies that addressed such questions. Indeed the different forms of assessment at each of the levels will be familiar to most readers. What makes our approach unique is the insistence on considering the often-overlooked assumptions about the nature of knowing and learning that underlie these different forms. These differences comprise the second feature of our approach.

### Increasingly Formal Representations of Domain Knowledge

The second feature of our approach is exemplified in the highly formalized characterization of knowledge in typical multiple-choice achievement tests, and the less formal characterization of that same content in an open-ended performance assessment.

Table 2: Three Views of Knowing, Learning, and Transfer, and their Implications for Assessment

	Rationalist		Socioculturalist
	Empirist		
Nature of Knowledge (epistemology)	Hierarchically organized associations that present an accurate but incomplete representation of the world. Assumes that the sum of the components of knowledge is the same as the whole. Because knowledge is accurately represented by components, one who demonstrates those components is presumed to know.	General and/or specific cognitive and conceptual structures, constructed by the mind and according to rational criteria. Essentially these are the higher-level structures that are constructed to assimilate new info to existing structure and as the structures accommodate more new info. Knowledge is represented by ability to solve new problems.	Distributed across people, communities, and physical environment. Represents culture of community that continues to create it. To know means to be attuned to the constraints and affordances of systems in which activity occurs. Knowledge is represented in the regularities of successful activity.
Nature of Learning (the process by which knowledge is increased or modified)	Forming and strengthening cognitive or S-R associations. Generation of knowledge by (1) exposure to pattern, (2) efficiently recognizing and responding to pattern (3) recognizing patterns in other contexts. "Higher-order" learning is the result of learning components.	Engaging in active process of making sense of ("rationalizing") the environment. The mind applying existing structure to new experience to rationalize it. One does not really learn the components of knowledge, so much as the higher-order structures needed to deal with those components later.	Increasing ability to participate in a particular community of practice. Initiation into the life of a group, strengthening ability to participate by becoming attuned to constraints and affordances. Having and being assigned an identity with the knowledge practices that define a community.
Manner in which Knowledge Transfers	Depends on the number and nature of individual associations needed in the new situation that were acquired in the previous situation.	Depends on the higher-level knowledge structures needed to solve problems in a new situation, relative to the structures that were constructed to solve problems in the previous situation.	Depends on the resources (physical and social constraints and affordances that scaffold participation) available in the new situation relative to the resources that supported participation in the previous situation.
Ideal Assessment (the ideal for assessing whether knowledge transfers)	Assess knowledge components. Focus on mastery of many components and fluency. For larger-scale assessments, use psychometrics to standardize.	Assess extended performance on new problems. Credit varieties of excellence.	Assess participation in inquiry and social practices of learning (e.g. portfolios, observations) Students should participate in assessment process. Assessments should be integrated into larger environment

We believe that systemic validity is maximized when we interrogate the assumptions about knowing and learning behind our assessment practices. These assumptions are often implicit, and proponents of particular assessment practices can be quite resistant to examining or even acknowledging them as assumptions. We further contend that the collective potential of multiple levels is maximized by emphasizing these differences. It is in this regard that our framework applies the sort of "competitive" approach to knowing and learning that is best exemplified in the analysis of Greeno, Collins, et al. (1996). In our experience, this is the most provocative aspect of our approach.

We contrast the three "grand theories" of knowing and learning using the widely accepted labels of *empiricist*, *rationalist*, and *socioculturalist*. Table 2 summarizes these three different views of knowing, learning, and transfer, and the ideal method for assessing that knowledge. Space considerations preclude a more extended discussion of all three perspectives. However, the implications for authentic assessment are detailed in Greeno, et al., (1996), Hickey and Pellegrino (2005), and Hickey and Zuiker (2003). A particularly provocative aspect of this comparative approach is its sharp delineation between contemporary sociocultural and views of knowing and learning and modern cognitive views. So called "situative" sociocultural views of knowing and learning reject the notion that knowledge is acquired by, and resident in, the minds of individual knowers (e.g., Greeno, et al., 1998, Wenger, 1998). Rather, knowledge is viewed as being constructed through, and fundamentally residing in, ritualized cultural practices. For us, the essential feature of these views is that they treat collective participation in social activity as the primary phenomenon in knowledgeable human activity. In doing so, they treat the cognition of individuals and the behavior of individuals as secondary phenomenon.

Sociocultural perspectives are certainly acknowledged in recent calls to broaden assessment practice (e.g., NRC, 2001a, 2001b). But most of these considerations take for granted individually-oriented cognitive accounts of knowing and learning. In doing so, they under-represent sociocultural perspectives as a way of understanding how social factors influence individual cognition (Hickey & Pellegrino, 2005). Arguably, these considerations are consistent with what Rogoff (1998) labeled *social influence* theories, and what Lave and Wenger (1991) called *cognition-plus*, to distinguish them from more uniquely sociocultural characterizations of knowledge.

As the assessment practices outlined below illustrate, we believe that the "discursive" representation of knowledge that is unique to sociocultural views is ideal for classroom assessment practices that are intended to directly advance student understanding. In this regard, our work draws very strongly from the considerations of assessment practice that start from a sociocultural perspective (e.g., Gee, 2003; Gipps, 1999). What is unique is the way we embed the refinement of sociocultural assessment practices within more conventional individually-oriented assessment practices. As shown in the seventh column of Table 1, we have found that the ideal assessment model is one that uses socioculturalist representations at the more immediate levels, rationalist representations at the middle levels, and empiricist representations at the more remote levels. The most important point is that the representations of knowledge become more *formal* as the distance of the assessment level increases. Specifically, we assume that the highly decontextualized representations of knowledge that appear necessary in external achievement tests are more

formal than the more context specific representations that are typical of proximal level assessments such as formal classroom assessments; these in turn are more formal than the contextualized discursive representations of knowledge in immediate and close-level assessment

At this point, it seems appropriate to introduce a critical caveat about our use of these perspectives in our framework. The distinction between these different characterizations of domain knowledge is a philosophical rather than scientific issue. This is not our debate. Our focus on sociocultural representations of knowledge at the immediate and close levels does not directly question the reality of more conventional individual representations of knowledge. More specifically, our focus on scaffolding collective participation at these levels does not deny that individuals are acquiring higher order schema and/or more specific lower-level associations via that activity. To the contrary, our guiding hypothesis is that focusing on directly collective participation at the immediate and close levels (and in practice, ignoring individual knowledge acquisition) will have a greater *indirect* impact on individual knowledge at the proximal level and beyond, compared to formative assessment practices that focus more directly on individual knowledge. We have yet to test this hypothesis formally, but are laying the groundwork for the complex research designs needed to do so. For now, we are focusing on assembling a convincing body of empirical evidence that our socioculturally-oriented classroom assessment has dramatic indirect impact on classroom performance assessments and statistically significant impact on external achievement tests.

### Crossing Knowledge Representations and Levels

Following is a detailed consideration with examples from two studies involving the GenScope software. These descriptions are intended to be informational and inspiring, rather than warranting empirical claims about the impact of specific formative assessment practices.

#### Immediate-Level Event-Oriented Observations

Directly following from the cultural characterization of knowledge summarized above, knowledge at this level in our approach is represented by the enactment of knowledge practices in specific curricular routines. Specifically, immediate-level assessments are concerned with collective participation in discourse practices of the domain as a particular lesson is enacted. By discourse, we mean conversations, as well as any interaction with the symbols and signs of the domain. Our actual assessment practices consist of informal observation during curricular routines. Because teachers and students can directly observe discourse, formative feedback can also directly and immediately enhance it. Because immediate-level assessment functions on a timescale of minutes, it is ideally suited for directly advancing "discursive" knowledge on a moment-to-moment basis. Meanwhile, the feedback from immediate-level assessment also provides teachers with useful guidance regarding the structure of the individual curricular routines, and guiding their refinement. As such, the formative priority of assessment at the immediate level should be the knowledge of the individual student at the time the assessment is

completed. Other functions, such as assigning student grades, are likely to undermine formative value for students.

In practice, our immediate level efforts consist mostly of guidelines for students and teachers that help ensure that productive domain discourse ensues when specific curricular routines are enacted. A specific inspiration in this regard was the criteria for guiding formation of student explanations tabled in Duschl and Gitomer (1997). In addition to trying out various enactments of such guidelines, we have also worked to revise curricular materials. In our initial GenScope project, the software developers had created worksheets for the guided inquiry investigations that allowed students to conduct investigations using the software's modeling capabilities. One of the investigators (Ann Kindfield, a science educator and geneticist) enhanced the student versions of the worksheets with densely detailed and technically accurate descriptions of biological phenomena. The teacher versions of the worksheets were enhanced with color coded "key points." Further, each activity targeted the use of technically accurate prose and diagrams. We encouraged teachers to promote their use by referring students to those descriptive explanations when conducting the investigations and reviewing them with the class after students completed the investigations. In other projects currently underway, we are starting to provide teachers with print-based and video-based guidelines that present specific examples of the forms of student discourse that should occur for very specific curricular routines. We also contend that an intensive focus on classroom discourse is particularly important for members of ethnic and linguistic minorities who have limited opportunities to participate in authentic academic discourse. These same academic forms of discourse embody the nuanced domain knowledge that is needed to solve challenging conceptual items on high-stake tests (e.g., Boaler, 1998) and succeed in more advanced coursework.

### Close-Level Activity-Oriented Quizzes and Learner-Oriented Feedback Rubrics

A great deal of our effort has been invested in refining close-level "quizzes" that are completed after perhaps four hours of curricular activity. We have written about these assessments elsewhere, but without actually labeling them as such (e.g., Hickey, Wolfe, & Kindfield, 2000). Unlike the immediate-level assessments, the quizzes require individuals to generate a written response to each item. The quizzes are "activity-oriented," in that their representation of domain knowledge is very similar to the representations in the curricular routines. These 'semi-formal' assessments build on the curricular representations around which students and teachers are presumed to have negotiated a shared understanding. This in turn provides the common knowledge base that is crucial for assessing and further advancing this understanding.

In the GenScope project, our quizzes used the fanciful "dragons" that teachers and students had become familiar with during the computer-based investigations. Students completed a quiz consisting of about five short-answer, open-ended items after roughly four hours of GenScope investigations. As detailed in Hickey, Kindfield, et al., (2003), our core innovation emerged in our efforts to maximize the formative value of these quizzes. After each quiz, students collaboratively review their completed assessments using a "learner-oriented" formative feedback rubric. These rubrics offer detailed explanations of the reasoning behind each quiz item, without directly stating the "correct" answer. Unlike

conventional scoring rubrics, they also include details that are not technically "necessary." Students use their completed assessments and the rubrics to discuss their understanding of the assessed topics during carefully-orchestrated "assessment conversations."

The first GenScope assessment project was designed in part to formally test the impact of different grading practices (i.e., norm-referenced vs. criterion-referenced). After the first cycle in the initial GenScope project, we asked teachers to stop assigning grades on the quizzes. Our reasoning highlights a critical insight for our work. When teachers assigned grades to student performance on the quizzes, the increased summative function clearly undermined our formative goals. Specifically, when students were assigned grades on the quizzes, attention and discourse during the feedback conversations shifted towards things like the "fairness" of the items and corrosive comparisons of prior performance. This undermined our goal of getting students to focus on and advance their fluency with the domain knowledge underlying the particular item. Furthermore, the process of grading consumed valuable time that was better spent modeling and coaching the feedback conversations.

A great deal of the design-based refinement described in the third section of this article is concerned with the appropriate level of formality for our close-level assessment. This means that we are constantly refining the formality of the knowledge representation on the quizzes (i.e., making it more or less similar to the curricular routines) and the formality of the assessment context (i.e., having teachers examine completed quizzes before giving students feedback). The quality of discourse during the feedback conversations and initial performance on the proximal level exams together provide useful evidence for guiding these refinements.

### Proximal-Level Curriculum-Oriented Exams

Proximal level assessments are concerned with the entire curriculum, and function on a timescale best characterized in terms of weeks. In the GenScope study, we created an open-ended performance assessment that was based on a comprehensive model of the development of reasoning in the domain of introductory genetics (Kindfield, Hickey, & Yessis, 1999). In other projects we have been assembling these exams by "cherry picking" items from released state tests that are aligned to the standards that our curriculum target. The exams are administered after an entire curriculum, and used by teachers to assign students grades for that part of the curriculum. This more formal representation and more summative administration is necessary for efficiently providing accurate feedback that is useful for both formative and summative goals. This includes (1) helping all students understand how domain knowledge is formally represented; (2) formally remediating specific topics or students in this regard; (3) revising the curriculum, and (4) fine-tuning the quizzes.

Most recently, we have begun providing students with learner-oriented formative feedback rubrics for the exams. This is a relatively new part of our approach, and one that we think is particularly promising. When exam items are carefully selected, the feedback rubrics can help students see how their new knowledge is (and is not) manifested in high-stake tests. Such exams can include released items that illustrate the logical "traps" or persistent misconceptions that item writers exploit to make difficult items for simple

concepts. The answer explanations direct attention to the domain concepts that the item writer had in mind, rather than just the five "more-correct and less-correct" associations that were created in the effort to efficiently assess knowledge of that concept. Interestingly, it appears that completing the close-level quizzes prepares students and teachers to take full advantage of the potential of exam conversations. In particular, it seems to minimize the corrosive effect that assigning grades normally has on feedback conversations.

Perhaps nowhere in our framework is the overall value of sociocultural perspective more apparent than in our understanding of the value of discursive formative feedback on proximal-level formal classroom assessments. Measurement specialists assume that items that exploit the nuances of domain language or student misconceptions are "bad" items that can and should be identified and removed using psychometric techniques such as differential item functioning (DIF). A sociocultural perspective allows one to understand this phenomenon quite differently. From this perspective, *all* individual assessments and tests are "peculiar but necessary" tools designed to serve a range of purposes (Hickey & Zuiker, 2003). Thus, the additional constraints on test format and representation necessary to accomplish the goals of increasingly distal level necessarily result in increasingly distorted representations of domain knowledge.

#### Distal-Level Criterion-Referenced Tests

Standardized multiple-choice tests that are aligned with specific content standards are the centerpiece of most test-based accountability programs. Such tests are usually assembled by selecting individual items from much larger pools of items that have been selected or written to be aligned with specific content standards. The formal administration context and the broad coverage of content make centrally-developed criterion-referenced tests problematic for evaluating the success of specific curricula. In our projects, we first create proxies for such tests and use them to refine and evaluate our curricula. For such proxy tests to provide valid and accurate estimates of distal-level impact, it is essential that they be developed and administered in a similar fashion as commercially developed tests. This requires some means of identifying the range of topics that will be covered. In the case of the GenScope project, we assembled a distal-level test using the roughly 75 genetics items that were included in released practice forms of the SAT II Biology Subject tests. Item difficulty was first used to rank the order of the pool of items. We then quasi-randomly sampled items by selecting every fifth item to create a pool of 15 items.. In current projects, we have scoured released high-stakes tests and test prep materials to locate items that are aligned with the content standards that the curriculum is intended to address.. We assemble pools of 5-10 items, rank them in order of apparent or stated difficulty, and then randomly select enough items to make a test that is of manageable length.<sup>4</sup>

One of our most important findings concerns the difficulty of obtaining significant gains on valid distal-level tests. By definition, such tests end up including a range of items that simply will not be covered by a specific intervention or curriculum. This sets a very high standard against which the success of iterative refinements of a curriculum (and in our case, associated assessments and formative feedback) can be evaluated. Indeed, we have found that scores do not increase at all or even go down slightly from pretest to posttest in the first implementation, while our non-implementation comparison classrooms typically

yield small increases (e.g., Hickey, Kruger, et. al., 2003). Such initially discouraging findings might lead instructional innovators to abandon their efforts or to not administer distal-level assessments. Because distal-level assessments are necessary for validly comparing multiple curricula, obtaining scores on such tests is essential for supporting continued use of the innovation. Rather, we think it makes more sense to not even administer distal-level assessments in the initial implementation cycles, and even to wait until acceptable performance is attained on the proximal level exams.

### Remote-Level Norm-Referenced Tests

Remote-level assessments provide evidence of change over time, operating on a timescale of years. Norm-referenced achievement tests are standardized in a way that allows an individual to be judged against a nationally normed sample, and in a manner that is stable from one year to the next. Sophisticated psychometric techniques are used to scale items to ensure that scores can be compared across forms. In the United States, the Iowa Test of Basic Skills (ITBS, Riverside Publishing, 2003) is one of several such tests that is widely used to measure changes over time in student achievement, and against a nationally-normed sample. A particularly useful feature of the ITBS and other such tests is the calibration of the students' score to the curricular week. This means that student percentile scores are adjusted to account for the actual week in the school year during which the test is completed. This provides more precise estimates of achievement when used in the kind of large-scale studies we envision; it also highlights that just one or two more correct responses on such tests can represent several months' worth of increased achievement. The National Assessment of Educational Progress (NAEP) is an even more remote-level test, and is used to measure the broader trends in student achievement over many years. There are numerous other remote-level outcomes that policy makers use (or might use) to judge changes in achievement over time, such as scores on college entrance exams, initial college performance, etc.

To measure increased achievement or attainment over time, remote-level assessments must be even more removed from a particular curriculum than distal-level assessments. While remote-level assessments provide feedback that is useful for evaluating the long-term impact of large-scale reform efforts, they are quite insensitive to fairly significant curricular reforms and innovations. As such, remote-level assessments are essential for assessing the consequences of efforts to maximize criterion-referenced distal-level performance. As mentioned in the introduction, a major controversy associated with the No Child Left Behind act in the US concerns declines on remote level outcomes (particularly the norm-referenced ITBS) as scores on the distal-level criterion-referenced tests targeted by the reforms were increasing. We aim to show the opposite in that our efforts to maximize distal-level outcomes have a measurably positive impact on corresponding remote-level outcomes.

While we have yet to measure remote-level outcomes, we look forward to doing so in our new elementary mathematics project. Highlighting the interpretive power of a multi-level model, differences in remote-level gains should "echo" gains on the distal-level tests. In other words, gains on norm-referenced tests due to the interventions should be associated with even larger gains on the criterion-referenced tests. This reveals whether the crucial

distal-level gains are merely the result of a narrowing of the curriculum to that test, or whether they are the result of success in more broadly improving teaching and learning. Ultimately, we hope to show that the inherent stability of remote-level assessments provides quasi-experimental evidence in support for a reform that is actually more scientifically valid than experimental evidence with distal-level assessments. Given the challenges of random assignment and comparison grouping in educational research, this seems like a particularly promising aspect of our approach.

### Increasingly Formal Cycles of Design Research

The third feature of our approach is the use of increasing formal cycles of design research. This feature of our approach is quite crucial, because the ideas outlined in the preceding section cannot be fully appreciated when isolated from these newer research methods in which they emerged. We have been particularly inspired by design-research theorists whose visions reflect a more contextualist worldview and a more sociocultural view of knowing and learning. These include Brown, (1992), Collins (e.g., 1999) Cobb, Confrey, DiSessa, Lehrer, and Schauble (2003) and Schoenfeld (in press). Design-based approaches emphasize the creation of "intermediate-level" practical theory, within iterative cycles of refinement.

Space limitations preclude further elaboration, but we believe our work is best understood in light of Stokes (1997) "use-inspired basic research" and the *engineering* approach outlined by Burkhart and Schoenfeld (2005). In order to create what Lagemann (2002) called "usable knowledge" we refine our insights about assessment in iterative cycles that aim to directly maximize the formative value of feedback at one level, while indirectly maximizing summative performance at the next level. This "engineering" of learning outcomes is done in close collaboration with teachers (*and* students) to make refinements that are *informal* (e.g., during class), *semi-formal* (e.g., from the first quiz to the second) and *formal* (from one year to the next).

### The Three Cycles in Our Framework

In our experience, it takes three cycles of design-based assessment around a particular innovation to maximize formative and summative potential across three levels of assessment. While we have obtained statistically significant gains on distal assessments in just two cycles (e.g., Taasobshirazi, Zuiker, & Hickey, 2005) we advise the use of the three cycles. Following is a brief summary of the three cycles that frame our efforts.

#### Implementation Cycle

The first cycle is partly akin to a conventional pilot study, but serves a very different purpose. This cycle requires intense collaboration with a small number of implementation teachers and students in order to refine the curriculum and the assessments. In most cases, a single class taught by a single teacher seems appropriate. Unlike conventional pilots that aim to identify contextual influences that need to be controlled for, design-based implementations search contextual variables so that they can be understood and

incorporated into the curricular innovation. A central part of this process is uncovering existing school and classroom culture that impacts assessment. For example, we have found that some classrooms have a particularly strong tendency to give students "points" for nearly every activity in which they are asked to participate. In these classrooms, we have found it necessary to continually fine tune the incentives attached to the quizzes and exams and the corresponding feedback conversations. In many cases our teachers have found it necessary to give students participation points to provide the initial motivation to engage in the feedback conversations; we are currently examining the extent to which this undermines the formative goals of that activity, and trying to better align the incentive value of external high-stake tests and exams to motivate engagement in feedback conversations.

The primary research method in this first cycle is discourse analysis (e.g., Gee & Green, 1998; especially Gee, 2003) and the primary target is the quality and quantity of domain discourse conversations that our assessment practices support. Video technology plays a central role in our work. Normally at least one group of consenting students is continuously recorded during all assessments and feedback conversations. While informal refinements are made "on the fly," the video recordings are scrutinized immediately to guide semi-formal refinements, and then later analyzed more intensively to guide formal refinements and provide naturalistic evidence of program impact. Theoretically, the assessments, the feedback materials, and our model for using them, are fundamental "participants" in that discourse, much as the students and teachers are.

### Experimentation Cycle

The second design research cycle is partly akin to conventional quasi-experimental classroom studies that are familiar to many educational researchers. This iteration is "experimental" because it aims to identify the relative contribution of different program components to different assessment outcomes. At this stage, additional implementation teachers may be added, and comparison classes may provide useful evidence (though only with assessments at the proximal level and beyond). Typically, we used within-teacher/between-class and between-teacher/within-school comparisons, and measured the relative and combined impact of the quizzes and exam.

One of the important advantages of the multi-level model is the way that numerous investigations can be embedded in even a single implementation. Space constraints preclude us from outlining even a fraction of the examples that we have attempted or considered. Some yield very specific insights, such as between-group/within class studies of the impact of alternative feedback rubrics on feedback conversations and exam performance. Others aim to build more generalizable knowledge such as between-class/within-teacher comparisons of the impact video-based discourse coaches. Because of the way that professional development constrains scalability, between-teacher/within-school comparisons of less-intensive and more-intensive profession development practices appears particularly important for large-scale success. The experimentation cycle of design-experimentation is a primary model of practice that appears most likely to deliver the most valued outcomes but that can be most readily implemented by other teachers and students who might be asked to use it.

## Evaluation Cycle

Our third cycle is partly akin to formal educational program evaluation. This cycle aims to develop rigorous evidence of impact on distal, and remote level assessment. While some situations seem likely to support truly random assignment, it appears that the most promising configuration is the identification of pairs of similar teachers or schools, and random assignment within pairs to implementation and comparison conditions. With multiple pairs of teachers or schools, stratified random assignment may be possible, affording the ultimate test of the overall reform. The presence of multiple levels of assessment affords a powerful interpretive framework for the distal and remote level scores that are generated in a large-scale evaluation. Experimental effects that fall shy of traditional criteria for being judged statistically unlikely (i.e.,  $p < .05$ ) take on additional meaning when considered in light of other theoretically-related outcomes at adjacent levels. This is critical for documenting continuous improvement of test scores for ethnic and linguistic subgroups (as required by *No Child Left Behind*) because it helps distinguish incremental success from random fluctuation.

## Prior Results and Future Directions

Our efforts so far have involved innovative science education environments. These are producing substantial evidence that our approach consistently raised high-stake achievement scores, while supporting the kind of inquiry-oriented learning environments supported by most science educators. The first GenScope project (launched in 1995 at the Educational Testing Service) provided initial insights and evidence regarding conversational feedback and multi-level assessment (Hickey, Wolfe, & Kindfield, 2000). The "enhanced" GenScope curriculum enabled many lower-achieving students to solve the kind of authentic inheritance problems that secondary students seldom master. In the fourth annual implementation cycle, the average gains across two classes on the proximal-level open-ended assessment were 3.07 standard deviations. This was much larger than the gain of 1.33 *SD* in two classes of comparison students whose teacher used a textbook (Hickey, Kindfield, et al., 2003).

A second cycle of GenScope research (Hickey, 2000) provided additional insights about the value of design-research methods along with our first evidence of distal-level impact. We carried out three annual refinements of a 20 hour-curriculum that featured a single level of discursive feedback on three close-level quizzes. In the four classes taught by our focal teacher, average gains on the proximal-level performance assessment increased from 0.65 *SD* in year one to 1.52 *SD* in year two, reaching 1.98 *SD* by the third year; similar students in two other matched classes using a conventional text-based curriculum had gained just .25 *SD*. On the distal-level test (the random sample of SAT II genetics items) average gains for the GenScope students increased from .21 to .74 to 1.06 *SD* across the three years; while the comparison students gained .57 *SD* (Hickey, Kruger, et al., 2003; Hickey, et al., 2004).

As mentioned above, a model with two levels of feedback conversations (close-level quizzes and proximal-level exams) have been deployed around three 20-hour multi-media science curriculum developed by the NASA-funded Center for Educational Technology

(Hickey, 2003). While funding constraints limited us to two annual implementation cycles, we nonetheless obtained statistically significant gains through the distal-level test on two of the three curricula (Taasobshirazi, et al., 2005). Another project launched in 2004 (Barab, Herring, Hickey, & Blanton, 2004) applies our assessment framework to *Quest Atlantis* (Barab, et al., in press), a multi-user virtual environment for middle school science.

One of the limitations of our studies so far is that we have lacked the resources to measure distal-level outcomes other than achievement tests, and have used an innovative curriculum that constrains the generalizability of our model and our conclusions. These issues are being addressed with new efforts around elementary mathematics. A pilot study was completed in two iterations with a 3rd grade fractions curriculum in Spring 2004 (Hickey, 2004), and a much larger study has just been funded by the U.S. National Science Foundation (Hickey, Mewborn, Beckmann, Lanehart, & Cohen, 2005). This new study should be quite informative, because we have the resources to (1) address an entire curricular year of fifth-grade mathematics, (2) implement with all fifth-grade teachers at one school on the third year, and (3) carefully measure distal-level outcomes that are consistent with all three views of knowing and learning. Specifically, in addition to district-administered criterion-referenced tests, we will assess distal-level understanding (by having students individually complete standards-oriented open-ended assessments) and distal-level participation in mathematical discourse (by analyzing the videotaped conversations of triads of students collaborative work on standards-oriented open-ended assessment). This will allow us to carefully measure the extent to which our enhanced participation in discourse and increased student understanding will transfer to new contexts. Finally, this study will also assess remote-level achievement outcomes. As such, we believe this new study will provide evidence supporting this approach that will be convincing to even the most skeptical observers, regardless of their theoretical orientation. Along the way, we expect to further expand the body of useful insights in a way that should inform assessment practices more broadly, and continue to help others contribute to this knowledge.

### Acknowledgements

This paper is based on work that was supported by the U.S. National Science Foundation Grant REC-0196225 to the University of Georgia and by the *Classroom of the Future Program* at the Center for Educational Technologies at Wheeling Jesuit University, which is funded by the U. S. National Aeronautics and Space Administration (NASA). The opinions presented here belong to the authors and do not necessarily represent the positions of the University of Georgia, the National Science Foundation, or NASA. The authors wish to thank Dr. Steven McGee, formerly of the Center for Educational Technologies, for his support and substantive contribution to some of the research described in this paper. Kate Anderson and Dionne Cross contributed substantively to the ideas in this article.

### References

- Anderson, L. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. Mahwah, NJ: Erlbaum.

- Barab, S.A., Herring, S., Hickey, D., & Blanton, B. (2004). *Quest Atlantis: Advancing a socially-responsive, meta-game for learning*. Grant REC-0411846 from the National Science Foundation to Indiana University.
- Barab, S.A., Thomas, M., Dodge, M., Carteaux, R., & Susun, H. (in press). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development*.
- Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, 5, (1), 7-74.
- Boaler, J. (1998). Alternative approaches to teaching, learning, and assessing mathematics. *Evaluation and Program Planning*, 21, 129-141.
- Brown, A.L. (1992). Design experiments: Theoretical and methodological challenges in creating complex interventions in classroom settings. *Journal of the Learning Sciences*, 2, 141-178.
- Burkhardt, H., & Schoenfeld, A.H. (2005). Improving educational research; toward a more useful, more influential, and better-funded enterprise. *Educational Researcher*, 32 (9), 3-14.
- Cobb, P., Confrey, J, DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32 (1), 9-13.
- Collins, A. (1999). The changing infrastructure of educational research. In E.C. Lagemann & L.B. Schulman (Eds.), *Issues in educational research. Problems and possibilities* (pp. 289-298). San Francisco: Jossey-Bass.
- Duschl, R. A. & Gitomer, D. H. (1997) Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment*, 4, 37-73.
- Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher*, 18 (9), 27-32.
- Gee, J. P. (2003). Opportunity to learn: A language-based perspective on assessment. *Assessment in Education*, 10 (1), 25-44.
- Gee, J. P. & Green, J. (1998). Discourse analysis, learning, and social practice: A methodological study. *Review of Research in Education*, 23, 119-69.
- Gipps, C. (1999). Sociocultural aspects of assessment. *Review of Research in Education* 24, 355-392.
- Greeno, J.G., Collins, A.M., & Resnick, L. (1996). Cognition and learning. In D. Berliner & R. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 15-46). New York: MacMillan.
- Greeno, J.G., (1998). The situativity of knowing, learning, and research and the middle school mathematics through application project. *American Psychologist*, 53, 5-26.
- Hickey, D.T. (2000). *Assessment, motivation, & epistemological reconciliation in a technology-supported learning environment*. Grant REC-0196225 from the National Science Foundation, Division on Research, Evaluation, & Communication to the University of Georgia.

Hickey, D.T. (2001). *Dimensions of participation in collaboration*. Grant from the NSF-funded Center for Interactive Learning Technologies Seed Grant Program to the University of Georgia.

Hickey, D.T. (2003). *Design-based implementation and evaluation of NASA CET multimedia science curriculum*. Subcontract from the NASA Center for Educational Technology to the Learning and Performance Support Laboratory.

Hickey, D.T. (2004). *Technology-supported multi-level assessment for improving mathematical teaching, learning, and achievement*. University of Georgia Faculty Research Grants Program.

Hickey, D.T., Kindfield, A.C.H., Horwitz, P., & Christie, M.A. (2003). Integrating curriculum, instruction, assessment, and evaluation in a technology-supported genetics environment. *American Educational Research Journal*, 40 (2) 495-538.

Hickey, D.T., Kruger, A.C., Fredrick, L.D., Schafer, N J., Russell, H.A., Bable, B., Hand, B., Michael, M., & Zuiker, S. (2003, April). *Design experimentation using multiple perspectives: The GenScope Assessment Project*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Hickey, D.T., Mewborn, D.S, Beckmann, S., Lanehart, S.L., & Cohen, A.S. (2005). *Multi-level assessment for enhancing mathematical discourse, curriculum, and achievement in diverse elementary school classrooms*. Grant REC-0440261 from the National Science Foundation's Research on Learning Environments (ROLE) program to the University of Georgia.

Hickey, D.T., & Pellegrino, J.W. (2005). Theory, level, and function: Three dimensions for understanding the connections between transfer and student assessment. In J.P. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective* (pp. 251-253). Greenwich, CT: Information Age Publishers.

Hickey, D.T., Wolfe, E.W., & Kindfield, A.C.H. (2000). Assessing learning in a technology-supported genetics environment: Evidential and consequential validity issues. *Educational Assessment*, 6 (3), 155-196.

Hickey, D.T., & Zuiker, S.J. (2003). A new perspective for evaluating innovative science learning environments. *Science Education*, 87 (3), 539-563.

Hickey, D.T., Zuiker, S.J., & Kindfield, A.C.H. (2004, April). *Curricular overview and learning outcomes in the GenScope Assessment Project*. Symposium presentation at the annual meeting of the American Educational Research Association, San Diego.

Hoff, D.J. (2004, March 10). Accountability conflicts vex schools. *Education Week*, 23 (26), 1, 23.

Horwitz, P., & Christie, M. (2000). Computer-based manipulatives for teaching scientific reasoning: An example. In M.J. Jacobson & R.B. Kozma, (Eds.), *Learning the sciences of the Twenty-first century: Theory, research, and the design of advanced technology learning environments* (pp. 163-191) Mahwah, NJ: Erlbaum.

Kennedy, M.M. (1999). Approximations to indicators of student outcomes. *Educational Evaluation and Policy Analysis*, 21, 345-363

Kindfield, A.C.H., Hickey, D.T., & Yessis, L. (1999, March). *Assessing student understanding of genetics: The NewWorm Assessment*. Paper presented at the Annual Meeting of the National Science Teacher's Association, Boston, MA.

Lagemann, E.C. (2002). *Usable knowledge in education*. Presidential memorandum. Spencer Foundation. Available online at: [http://www.spencer.org/publications/usable\\_knowledge\\_report\\_ecl\\_a.htm](http://www.spencer.org/publications/usable_knowledge_report_ecl_a.htm)

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge: Cambridge University Press.

Lemke, J.J. (2000). Across the scale of time: Artifacts, activities, and meaning in ecosocial systems. *Mind, Culture, and Activity* 7 (4), 273-290.

Markley, M. (2004, June 6). TAAS scores rose as SATs fell: Some say states focus on basics comes at the expense of college prep. *Houston Chronicle* [Accessed online at <http://www.chron.com/cs/CDA/ssistory.mpl/metropolitan/2610999>].

National Research Council (2001a). *Classroom assessment and the national science education standards*. J. M. Atkin., P. Black, & J. Coffey, (Eds.). Washington, DC: Author. Available online at: <http://www.nap.edu/catalog/9847.html>

National Research Council (2001b). *Knowing what students know: The science and design of educational assessment*. J. W., Pellegrino, N. Chudowski, N., & R. W. Glaser, R. (Eds.). Washington, DC: National Academy Press. Available online at: <http://www.nap.edu/catalog/10019.html>

National Research Council (2003). *Bridging the gap between large-scale and classroom assessment: Workshop report*. Committee on Assessment in Support of Instruction and Learning. J. M., Atkin, Chair. Available online at: [http://www7.nationalacademies.org/bota/Bridging\\_the\\_Gap.html](http://www7.nationalacademies.org/bota/Bridging_the_Gap.html)

Popham, J. W. (2003). The seductive allure of data. *Educational Leadership*, 60 (5), 48-51.

Riverside Publishing (2003). *Iowa tests of basic skills, Form A*. Product details from publisher website at: [http://www.riverpub.com/products/group/itbs\\_a/home.html](http://www.riverpub.com/products/group/itbs_a/home.html)

Rogoff, B. (1998). Cognition as a collaborative process. In D. Kuhn & R.S. Siegler (Eds.), *Cognition, perception and language (Vol. 2, Handbook of child psychology (5th ed.)*, W. Damon (Ed.) pp. 679-744). NY: Wiley.

Ruiz-Primo, M.A., Shavelson, R.J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, 39, 369-393.

Schemo, D.J., & Fessenden, F. (2003, December 3). *Gains in Houston schools: How real are they?* New York Times.

Schoenfeld, A.H. (in press). Design experiments. In P.B. Elmore, G. Camilli, & J. Green (Eds.), *Complementary methods for research in education*. Washington, DC: American Educational Research Association.

Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational Researcher*, 29 (7), 4-14.

Sleek (2004). [Website offering test prep programs for various state tests]: <http://www.sleek.com/index.htm>

Sloane, F.C., & Kelly A.E. (2003). Issues in high-stakes testing programs. *Theory into Practice*, 42 (1), 12-17.

Stokes, D.E. (1997). *Pasteur's quadrant: Basic science and technical innovation*. Washington, DC: Brookings Institution Press.

StudyIsland (2004). [Website offering practice tests for 13 state tests]: <http://www.studyisland.com>

Taasoobshirazi, G, Zuiker, S.J., & Hickey, D.T. (2005). *Design-based implementation and refinement of an inquiry-orientated multimedia curriculum, assessments, and learning environment*. Presentation at the annual meeting of the American Educational Research Association, Montreal.

Tingley, B., Thrall, A., & Ward, G. (2001). High stakes management™: Leveraging *SuccessMaker*® research to raise achievement. Pearson Educational Technologies Available on line at: <http://www.pearsonedtech.com>

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge University Press.

Wilson, M. (2004) (Ed.). *Towards coherence between classroom assessment and accountability*. The 103<sup>rd</sup> Yearbook of the National Society for the Study of Education. Chicago IL: University of Chicago Press.

Ysseldyke, J., & Tardrew, S. (2002). *Differentiating math instruction: A large-scale study of accelerated math, first report*. Madison, WI: Renaissance Learning, Inc.,: <http://research.renlearn.com/research/129.asp>

## Notes

1. Nonetheless, test prep systems “succeed” because due to low per-student cost and self-paced individualized format, it is possible for modestly-funded studies of these programs to demonstrate statistically significant impact on “external” tests in randomized experimental studies—which is both necessary and sufficient to obtain federal education funding in the US.
2. Consider, for example, the inflation of experimental effects in one unpublished (but widely cited and distributed) external evaluation commissioned by one of the firms. Ysseldyke and Tardrew (2002) eliminated a sizable proportion of students from the experimental group who failed to complete a specified number of practice tests. Because such students are likely to be among the lowest performers, eliminating them from the experimental group (but not the control group) could have easily produced the experimental effect of the magnitude that was reported. We emphasize that the following should not be read as an effort to advance one theory of assessment over another. Rather, we have found these distinctions offer a useful way of understanding different views of transfer in order to better address practical issues in assessment, testing, and evaluation that follow from those differences.

3. In this regard, our characterization of immediate-level assessment diverges from that of Ruiz-Primo et al. (2002). Their immediate-level assessments examined the artifacts from the enactment of the curriculum, rather than the actual enactment.
4. We acknowledge that extracting items from complete tests and reassembling them in this fashion compromises their evidential validity. In selecting items based on stated difficulty or apparently difficulty, we are attempting to build a high-stake proxy that captures the entire range of items that might appear in a distal-level assessment. Concerns about the parameters of a collection of items are highly relevant to the evidential validity when trying to reliably sort the knowledge of individual students. But doing so is a very time-intensive and costly process, and well beyond the scope of what we are trying to accomplish in our framework. In our new project, once we have refined parts or all of a curriculum using proxy tests (in the initial implementation cycles), we would then evaluate a large-scale implementation of the curriculum using more conventional distal-level assessments, such as formally administered criterion-referenced tests.

### The Authors

DANIEL T. HICKEY is an associate professor in the Learning Sciences Program, Indiana University, USA. He is studying how new situative theories of learning and new models of classroom assessment can simultaneously improve classroom discourse, student understanding and high-stakes achievement. He also is studying new situative models of motivation and exploring the impact of assessment and testing on student engagement and motivation more broadly.

STEVEN J. ZUIKER is a doctoral student in the Learning Sciences Program at Indiana University, USA. He studies the relationship between formative and summative functions of assessment from a situative perspective. His current work considers the embedment of formative assessment practices in multi-user virtual environments.

GITA TAASOOBSHIRAZI is a doctoral student in the Applied Cognition and Development program at the University of Georgia, USA.

NANCY JO SCHAFER is an instructor in the Department of Early Childhood Education at Georgia State University, USA. She studies socio-cultural approaches to teacher preparation and the development of reflective practice with in a "Community of Learners" to improve teachers' pedagogy. Her primary focus is on classroom management in urban settings.

MARINA A. MICHAEL is a doctoral student in the Combined Program in Cognitive, Developmental and Educational Psychology at the University of Cyprus. She is interested in the various manifestations of students' motivation in educational and learning contexts. She studies "engagement" from different epistemological perspectives. Her current work considers engagement in relation to students' epistemology in elementary school settings.

Correspondence: <dthickey@indiana.edu>