

CHAPTER 11

Situative Approaches to Student Assessment: Contextualizing Evidence to Transform Practice

DANIEL T. HICKEY AND KATE T. ANDERSON

This volume of the *NSSE Yearbook* is concerned with using expanded types of evidence in order to understand and improve teaching and learning. Doing so presents the challenge of using evidence from one sector of the educational system to guide decision making in other sectors. These sectors are defined by a range of stakeholders, including policymakers, researchers, administrators, educators, and students themselves. Most of the actual evidence comes from assessments of student learning and achievement. Hence, a primary challenge is assessing students' knowledge in ways that result in evidence that is interpretable and usable by decision makers at different levels, which in turn raises the challenge of gathering useful evidence about the consequences of those decisions.

This chapter aims to introduce several ideas about using evidence from assessment to guide educational decision making. We expect these ideas to be new to many readers, as they reflect the influence of “socio-cultural” theories of learning (e.g., Vygotsky, 1986), particularly the theories of “situative” sociocultural theorists (e.g., Greeno & MMAP, 1998). These theories assume that all learning is *social* change. This contrasts with traditional theories underlying most prior considerations of assessment, which assume that learning is fundamentally about *individual* (“cognitive”) change. From a sociocultural perspective, the knowledge that students learn is distributed across the many diverse participants that contribute to every educational encounter. Therefore,

Daniel T. Hickey is an Associate Professor in the Indiana University Learning Sciences Program, and studies transfer of learning, assessment, motivation, and design-research methodologies. Kate T. Anderson is an Assistant Professor at the National Institute of Education's Learning Sciences and Technologies Group in Singapore. Her research examines participatory forms of discourse, identity construction, and social practice in the cultural and historical contexts of classrooms and schools.

“to learn” means to participate more successfully in the collective practices that define particular ways of knowing as recognized by various communities.

In recent years, several socioculturally oriented theorists have turned their attention to assessment (e.g., Beach, 2003; Gee, 2003). These considerations acknowledge the difficult task of operationalizing sociocultural views of learning in communities using practices that have long assumed individually oriented views of knowing and learning (see Haertel & Greeno, 2003). For example, many of the current controversies over student assessment (including those discussed in this volume) concern *which* individually oriented view of learning yields the most useful evidence for various decision makers. Nonetheless, we contend that sociocultural approaches have unique potential for accomplishing widely held goals for the improvement of educational research and practice.

In this chapter, we highlight the value of socioculturally inspired assessment practices using what we call *discursive* classroom assessment. In contrast to the individually oriented classroom assessments used by most teachers, discursive classroom assessments begin with the collective knowledge represented in classroom discourse, rather than students' individual conceptualizations. While these practices were developed across several multiyear projects involving innovative technology-supported science curricula, we also discuss how they can be used in other educational settings. We describe how these discursive assessments emerged in an initial effort to enhance learning gains in a 20-hour genetics curriculum, as measured by students' performance on a comprehensive assessment of their understanding of key concepts in inheritance. We then summarize the evidence from a subsequent effort to refine the discursive assessments in order to increase those gains and validate them against a comparison group using an “external” measure that predicted achievement on high-stakes criterion-referenced tests.

We expect that our discursive assessment practices and our insights about using, refining, and validating them should be relevant to educators and innovators who want to maximize the impact of specific curricula on classroom discourse in order to enhance students' understanding and achievement. They should also be relevant to readers who are interested in strategies for raising student achievement without resorting to “test-prep” methods. As will be shown, using assessments to scaffold participation in forms of discourse that indirectly (but consistently) raise achievement scores supports a very different

experience than being trained to improve performance on a specific test.

We present these examples retrospectively within a comprehensive assessment framework that emerged across subsequent projects. This framework introduces several socioculturally inspired notions that we believe may help obtain evidence from multiple types of assessment that are useful for improving teaching and learning. This framework also highlights the unique value of socioculturally oriented research methods for refining assessments. We illustrate how these methods can be used to improve the value of assessments for teaching and learning while minimizing the potential negative consequences of assessment (such as diminished student motivation or narrowed curricula). It is expected that this framework and associated methods will be particularly relevant to readers who are interested in “balancing” competing uses of different types of evidence. This includes educators who want to use classroom assessments to help students learn by virtue of completing them while also refining curricula and assigning grades, and administrators who want to provide evidence of achievement demanded by policymakers in ways that are most likely to support educational improvement. In this regard, we hope to show that our discursive assessments and use of design-based research methods represent promising extensions to the many well-established strategies for aligning curricula to external standards and tests (e.g., Wiggins & McTigue, 2005).

We then conclude by suggesting that sociocultural perspectives also have the potential for making sense of and addressing two broader challenges facing evidence-based educational reform: the controversies over competing individually oriented approaches to assessment, and the questionable validity of students’ gains on assessments that are directly targeted by reforms as evidence of more systemic improvement. It is hoped that readers interested in broader, long-term advancement in the use of evidence for educational improvement will find these suggestions thought-provoking and worthy of further consideration and debate.

A Discursive Approach to Classroom Assessment

Our framework for classroom assessment emerged in studies using *GenScope*, a computer-based modeling program developed for teaching introductory genetics in high school life science classrooms (Horwitz & Christie, 2000). This program features windows that correspond to

the various levels of biological organization, including *DNA*, *chromosome*, *meiosis*, *organism*, *pedigree*, and *population*. Each window features novel, interactive representations of genetic information and easy to use tools for manipulating that information. The software employs simplified organisms, primarily the “Dragons” shown in Figure 1. GenScope’s developers created a 20-hour curriculum incorporating inquiry-oriented activities where students explored concepts introduced in the program (e.g., dominance, sex-linked inheritance) by using the software to solve problems that involved more than one level of biological organization.¹

The first “GenScope Assessment Study” was initiated at Educational Testing Service to develop assessments that could be used to evaluate the 20-hour GenScope curriculum. The resulting *NewWorm* assessment included short-answer items and open-ended problems, using a different simplified organism and more conventional representations of genetic information than the software program itself did (Hickey, Wolfe, & Kindfield, 2000). This included “cause-to-effect problems” such as using the parents’ genotype in the familiar Punnett square in order to predict offspring genotype. It also included “effect-to-cause” problems where information about offspring is used to predict the parents’ genotype, as shown in Figure 2. This kind of reasoning is essential for geneticists and is seldom mastered by secondary students. In the 1996 National Assessment of Educational Progress (NAEP), less than 25% of secondary students were able to solve a multiple-choice effect-to-cause item akin to the one in Figure 2. The *NewWorm* assessment was intended for use as a “far-transfer” measure that was independent of any particular genetics curriculum, including GenScope. In other words, it was designed for use in classrooms regardless of the curriculum they were using to teach inheritance.

FIGURE 1
ORGANISM AND PEDIGREE WINDOWS IN GENSCOPE SOFTWARE

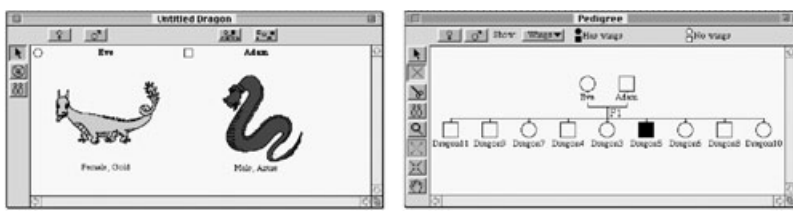


FIGURE 2

EXAMPLE NEWWORM ITEM ASSESSING EFFECT-TO-CAUSE REASONING

Another inherited characteristic in the NewWorm is Eyelids. Both NewWorm1 and NewWorm2 have clear eyelids. However when you mate them and produce 100 offspring, you find:

- 74 (51 males and 23 females) have clear eyelids
- 26 (0 males and 26 females) have cloudy eyelids

Remember: Males are XX and females are XY.

1. There are two alleles for Eyelids. Is the relationship between the two alleles simple dominance or incomplete dominance? Answer: _____
 - 1a. What is it about the *offspring* that indicates simple or incomplete dominance?

Student performance on the NewWorm assessment in the initial GenScope implementations was disappointing. Many students who successfully completed the various GenScope activities were unable to solve corresponding paper and pencil problems on the NewWorm assessment. Specifically, most of these students failed to solve the cause-to-effect NewWorm problems and none of the students could solve the effect-to-cause problems. Initially, it was unclear whether the students were failing to learn the underlying inheritance concepts in the activities, or whether they were gaining useful knowledge that failed to transfer to the new organisms and new representations in the NewWorm assessment context.

To resolve this question, the researchers developed a “near-transfer” problem-solving assessment that used the actual organisms and traits from the GenScope software targeting similar problems as the NewWorm. As reported in Hickey, Kindfield, Horwitz, and Christie (2003), the near-transfer assessment yielded better results, but they were still quite disappointing. This suggested that students were not learning key concepts and skills in the existing activities, as they were unable to reproduce this knowledge in a context that was relatively similar to the one in which they initially were expected to learn it. While these results convinced the curriculum development team to further refine the GenScope activities, it was also apparent that the near-transfer assessment had untapped “formative” potential for supporting further student learning. In other words, in addition to providing “summative” evidence about student learning to guide refinements of the curriculum, this new assessment had potential value for further advancing student understanding. For example, it provided information for students and

teachers to use in order to address student learning and misconceptions after completion. Initially, solving the problems on the near-transfer assessment was itself a learning opportunity because the problems were essentially extensions of the GenScope activities. It seemed that doing so, and then further reviewing solutions, might be a good way for students to develop the robust understanding needed to solve complex inheritance problems.

The various problems on the near-transfer assessment were then refined and organized into four *Dragon Investigations*. These investigations were used to reorganize the existing GenScope activities into four separate 5-hour units. Each investigation focused on one of four types of increasingly challenging problems, targeting increasingly challenging concepts. A scoring rubric for teachers was developed for each investigation. The implementation teachers were encouraged to have students complete the investigation at the end of each unit and score the completed investigations, and then provide students with formative feedback to further their understanding.

Our efforts to enhance the formative value of the Dragon Investigations were inspired by Duschl and Gitomer's (1997) success at using portfolio assessments and "assessment conversations" to support scientific argumentation and discourse. To this end, new rubrics were developed for each Dragon Investigation; in contrast to the scoring rubrics (designed for efficient evaluation of students' responses by knowledgeable adults), these "answer explanation" rubrics were designed to foster student discourse and support additional learning. The answer explanations used relatively advanced prose and diagrams to explain the reasoning behind each item without directly stating the "correct" answer. Their readability was well above the grade level of the targeted students and introduced new content and technical terms that were not technically needed to "answer" the problem. Reflecting our nascent appreciation of sociocultural theories of learning, this was intended to help students situate their new understanding in broader contexts of use. This included other specific problem-solving contexts (such as the NewWorm) as well as subsequent life science courses.

Implementation teachers began using these answer explanations to facilitate what were called "feedback conversations." Across several refinement cycles, the team fostered increasingly sophisticated classroom discourse (and teachers' understanding of how to promote it) around challenging inheritance problems on the Dragon Investigations. They did so by exploiting (1) the class's collective familiarity with GenScope's organisms and representations of genetic information; (2)

students' motivation to determine whether they solved problems correctly; and (3) the scaffolding of the answer explanations. At the end of the first GenScope assessment project, students in most of the observed GenScope classrooms were routinely engaging in relatively sophisticated argumentation about inheritance.

Given the difficulty of fostering worthwhile discourse and argumentation in science classrooms (Duschl & Gitomer, 1997), the project's increased success in this regard was encouraging. Even more encouraging were the corresponding increases in gains on the far-transfer NewWorm assessment. Unlike the earlier implementations, many students successfully solved some or all of the challenging effect-to-cause problems. The final round of implementations yielded a 3.1 standard deviation (SD) gain on the NewWorm in a class that participated in feedback conversations. This was significantly greater than the 2.2 SD gains in another class taught by the same teacher who used a more conventional review of the Dragon Investigations. Meanwhile, similar students in the same school whose teacher used the existing textbook curriculum gained just 1.3 SD (Hickey et al., 2003). Given that most secondary biology students seldom achieve the level of reasoning attained by most of the GenScope students, this was encouraging evidence about the value of our emerging efforts to use new forms of assessment to improve students' understanding and achievement.

An obvious issue in this first study was that the introduction of the Dragon Investigation and associated formative feedback compromised the validity of student performance on NewWorm assessment as evidence of "far" transfer. In other words, the NewWorm's close alignment with the content and representations in the Dragon Investigations compromised the NewWorm's validity in cross-curricular comparisons with non-GenScope classrooms. This concern was addressed in a second multiyear effort to further refine and validate this new "discursive" approach to formative feedback (Hickey, 2001). For this project, an additional assessment was developed using a stratified random sample of released genetics items from the SAT II Biology Subject Area test and the NAEP science assessments (including the aforementioned NAEP item). As these new items were broadly aligned to the relevant state science standards and were entirely independent of any curriculum, the assessment offered a valid proxy for criterion-referenced achievement tests designed to target those same standards. This included the federally mandated achievement test, as well as the science subtest of the high school graduation test that participating students had to pass in order to receive their diploma.

Now that we have described the initial research context in which this approach was first developed, we will introduce the broader theoretical framework which emerged around these efforts before discussing further refinements of the Dragon Investigation discursive assessments and subsequent performance on the NewWorm and “proxy” achievement tests.

A Sociocultural Assessment Framework

In proposing and implementing the second GenScope Assessment Project, several key insights emerged and a comprehensive framework for evidence-based reform began to take shape. These insights and the nature of the framework reflected the influence of sociocultural views of knowing and learning, which were beginning to be widely appreciated at the time (e.g., Greeno & MMAP, 1998; Wenger, 1998). In particular, the following characterization of the “ideal” functions of different types of assessments drew from the comparative characterization of sociocultural views outlined in Greeno, Collins, and Resnick (1996; also Case, 1996).

Multiple Levels of Assessment

The most important feature of the new assessment framework was the identification of multiple assessment “levels,” characterized, in part, by how close their representations of content are to curricula (i.e., “distance”). The research team began characterizing the three types of assessments in the project using the levels defined in a summative evaluation conducted by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002): *immediate*, *close*, *proximal*, *distal*, and *remote*. While our characterization shared the underlying continuum of increased “distance” from a given curricular routine, the levels were reconceptualized. In the original characterization, immediate-level evidence was obtained by analyzing the artifacts that students generated in science classrooms (reports, worksheets, etc). As shown in Table 1, immediate-level assessment in the revised characterization concerns the collective discourse that takes place when specific curricular activities are enacted. In our case, this consisted of the informal observations that the researchers and teachers had been making while students worked together to complete the inquiry-oriented activities using the GenScope software.

Close-level assessment in our framework concerns students’ conceptual familiarity and participation in relevant discourse *after* a particular

TABLE 1
MULTILEVEL ASSESSMENT FRAMEWORK AND IDEAL FORMATIVE FUNCTIONS

Level	Orientation	Targeted Content	Relative Time Frame	Ideal Assessments	Appropriate Formative Function for Students	Ideal Formative Functions for Others
Immediate	Specific Curricular <i>Event</i>	Discourse during enactment	Minutes	Event-oriented observations (informal activity's enactment)	Discourse during the enactment of a particular activity	Teacher: Refining discourse during the enactment of a particular activity
Close	Specific Curricular <i>Activity</i>	Discourse and understanding associated with activity	Days	Activity-oriented investigations (semi-formal classroom assessments)	Discourse and understanding following the enactment of a particular activity	Teacher: Refining the specific curricular routines and providing informal remediation to students
Proximal	Entire <i>Curriculum</i>	Concepts taught in the curriculum	Weeks	Curriculum-oriented exams (formal classroom assessments)	Understanding of primary concepts targeted in curriculum	Teacher/curriculum developer: Providing formal remediation and formally refining curricula
Distal	<i>Standards</i> (targeted by a specific curriculum)	Regional or national content standards	Months	Criterion-referenced tests (external tests aligned to content standards)		Administrators: Selection of curricula that have the largest impact on achievement in broad content domains
Remote	<i>Achievement</i>	National achievement	Years	Norm-referenced tests (external tests standardized across years, such as Iowa Test of Basic Skills and National Assessment of Educational Progress)		Policy makers: Long-term impact of policies on broad achievement targets

activity is completed (as supported by our near-transfer Dragon Investigations). In contrast, *proximal-level* assessments concerns the broader conceptual understanding targeted by an entire curriculum (as in our far-transfer NewWorm). This differed from *distal-level* assessments that concern student achievement on the targeted content standards (as in our proxy achievement test). Finally, *remote-level* assessments concerned achievement gains relative to broader populations from one year to the next (as measured by norm-referenced tests, which has yet to be considered in the project).

At this stage, the research team also began using the notions of *orientation* and *timescale* to understand and characterize the difference between assessment levels. As shown in Table 1, each of the five assessment levels is oriented towards an increasingly broader characterization of activity: *events*, *activities*, *curriculum*, *standards*, and *achievement*, respectively. Lemke's (2000) notion of *timescale* helped further distinguish each level's formative potential by framing the temporal context in which assessment evidence is most relevant and useful (Zuiker, Hickey, Kwon, Chapman, & Barab, 2005). As shown in Table 1, the relative timescales associated with each level are *minutes*, *days*, *weeks*, *months*, and *years*, respectively.

This notion of assessment "distance" (i.e., levels), along with orientation and timescale, define three distinct theoretical continua. As elaborated in Hickey, Zuiker, Taasoobshirazi, Schafer, and Michael (2006), it is assumed that the formative value of a particular assessment is maximized when it defines a discrete location along the continua and when the appropriate orientation and corresponding timescale for that location are identified. In other words, insights from three different theoretical continua are combined in order to define one assessment's ideal function(s). For example, the formative potential of *immediate-level/event-oriented* observations of discourse during a specific event is greatest while the activity is still being enacted—a timescale of minutes. In contrast, *close-level/activity-oriented* assessments like the Dragon Investigations are completed *after* the activity is completed. As such, their formative potential operates on a longer timescale, roughly corresponding to days. This is different still from *proximal-level/curriculum-oriented* assessments like the NewWorm that are completed following the entire curriculum. This longer timescale makes them more useful for refining curricula and guiding formal review and remediation, but less useful for directly supporting student learning. Finally, the much longer timescales of *distal-level/standards-oriented* criterion-referenced tests and *remote-level/achievement-oriented* norm-referenced tests render

them nearly useless for directly supporting student learning, but highlights their respective value for evaluating both the impact of different curricula and the impact of different policies on student achievement. Table 1 summarizes additional insights about the ideal formative potential of assessments at each level. More details are included in Hickey et al. (2006) and additional examples are included in the succeeding sections.

Another key notion that emerged in this effort was a “unidirectional” assumption about transfer of learning from formative feedback. As elaborated in Hickey and Pellegrino (2005), this means that learning around more familiar representations of content knowledge (i.e., feedback on our close-level Dragon Investigations) will transfer more readily to semi-familiar representations of that knowledge (i.e., solving corresponding problems on our proximal-level NewWorm) than its converse. Likewise, proximal-level formative feedback (i.e., remediation and review based on the NewWorm) should transfer more readily to more abstract, unfamiliar representations of that knowledge (i.e., the corresponding distal-level achievement test items).

This assumption about knowledge transfer has yet to be directly supported by empirical evidence (and may not be amenable to generalizable proof). Nonetheless, it has lent itself to a coherent organizing framework for impacting performance on distal-level measures of achievement across a number of studies (as elaborated in the succeeding discussions). This assumption underlies our core strategy for evidence-based reform. In essence, (1) existing activities are organized around close-level/activity-oriented classroom assessments and discursive formative feedback rubrics with an eye toward targeted content standards. These are then (2) used to aid students’ participation in specified forms of collective discourse, in order to (3) increase performance on proximal-level/standards-oriented classroom assessment through expanded ways to understand and make sense of subject matter. Subsequently, (4) more conventional formative feedback practices around the proximal-level assessments are used to advance individual students’ understanding by supporting remediation, addressing misconceptions, and refining curriculum, in order to (5) increase the number of students meeting criteria on high-stakes achievement tests. Ideally, such an effort is also (6) evaluated by research that documents corresponding gains on norm-referenced achievement tests. The next section describes the research methods that make this seemingly unwieldy process surprisingly coherent and manageable.

Iterative Refinement of Formative Assessment

Our efforts to increase student learning via formative feedback build on design-based research methods (e.g., Barab & Squire, 2004; Cobb, Confrey, DiSessa, Lehrer, & Schauble, 2003). Before formally testing whether or not formative feedback helps enhance achievement (i.e., at the distal level), we iteratively refine the formative functions of the various levels. Essentially, we “engineer” (Burkhardt & Schoenfeld, 2003) the formative function of assessments at one level to help maximize initial performance at the next level. In the case of the close-level assessments, this means attempting different strategies for enhancing discourse during the close-level feedback conversations and then searching for evidence of those improvements in problem-solving performance on the proximal-level NewWorm assessment.

One of the most useful ideas regarding this iterative refinement is the value of increasingly formal design cycles within particular projects, where the insights, practices, and accountability associated with each cycle are incorporated into subsequent cycles. This aspect of our approach was first formally defined in a subsequent project (Hickey, 2003) involving three multimedia science curricula developed by the NASA-sponsored *Classroom of the Future* program. The overall framework is currently being refined in studies involving the *Quest Atlantis* multiuser virtual environment for elementary and middle school students (Barab, Herring, Hickey, & Blanton, 2004). It is also being refined and more formally evaluated in the context of its application to the entire fifth-grade *Everyday Mathematics* curriculum (Hickey, Mewborn, & Lewison, 2005).

While the iterative refinement framework was not formalized at the outset of the second GenScope project, the three annual GenScope implementations did follow roughly the same cycles. The following description includes examples and findings from all three projects, but draws most strongly from GenScope, with particular attention to the initial implementation cycle, because it focuses most strongly on innovative discursive assessments and related tools for scaffolding discourse.

Implementation cycle. The first cycle focuses on curricular activities and close-level assessments with the goal of promoting discursive practices and maximizing students’ performance on the proximal-level assessments (e.g., the NewWorm assessment). The research team works intensively with one or two teachers and relies mostly on discourse analytic methods in order to understand and then help teachers and

students engage in more meaningful collective participation in domain-specific discourse. Discourse analysis is purposeful, theoretically informed examination of who says what, how, to whom, and for what purpose. Our efforts draw strongly from the research literature on classroom discourse in general (e.g., Gee, 2001) as well as more focused consideration of discourse in the particular domain in which the authors are working (e.g., Jiménez-Aleixandre, Rodríguez, & Duschl, 2000; O'Connor, 2001).

A major part of the implementation cycle is ensuring that clusters of curricular activities are aligned to targeted content standards, and then creating close-level assessments and corresponding answer explanations. It has been found that this relatively informal process of alignment provides useful guidance as to how the specific activities should be enacted. In other words, this framework prompts curriculum designers and teachers to carefully consider the discourse that should emerge *during* the activity when designing the close-level assessments and the subsequent discourse that should be possible *after* the activity has been completed. The process and resulting materials help teachers and researchers to adjust the enactment of curricular activities “on the fly” to prepare students to participate successfully in the corresponding close-level investigation.

Many of these refinements are based on teachers' suggestions as we work with them quite closely during the initial implementation cycle. The close-level assessments provide teachers with useful evidence for shaping their own refinements to practice, and teachers are encouraged both to review the close-level assessment and answer explanations before enacting the targeted activities and to refine their enactments continuously. As relationships between teachers and the research team develop, teachers readily provide suggestions that are helpful in their own class and which often generalize to other classes (and projects) as well. Our goal for teachers is to initiate ways to scaffold discourse or to “take over” from the researchers. Teachers know their students better and have far more experience working with their curriculum's goals.

Videotaping classroom interaction has played a central role as evidence for refining discourse practices and for beginning to validate the impact of various refinements. Discourse analysis of videotaped feedback conversations around one of the NASA science curricula, for example, revealed how subtle aspects of teachers' strategies for engaging students in feedback conversations (e.g., “inserting” specific content when conversations began to falter) and students' ways of engaging the

topic with each other (e.g., stating answers and moving on versus questioning each other's reasoning) led to dramatic differences in students' participation and discourse (Anderson, Zuiker, Taasobshirazi & Hickey, in press). As the implementation cycle progresses, the focus moves toward ensuring that collective discourse around the close-level assessments does indeed support individual performance on the proximal-level assessments.

A range of design strategies is available for attempting different refinements and searching for evidence of improvement. For example, when teachers are working with more than one classroom, comparisons from one class period to the next can be used to informally refine some features, such as student grouping (whole class versus small group, homogeneous versus heterogeneous groups, etc.). Modifications from one close-level assessment to the next can also be carried out and examined, through, for example, changes to item formats (e.g., open-ended versus multiple choice) or the explicitness of the answer explanations.

Another focus of the implementation cycle is on the various tools that are developed for scaffolding discourse around close-level assessments. One strategy explored in the GenScope project was showing students video clips of themselves and their classmates that illustrated the features of good feedback conversations (Schafer, Kruger, Hickey, & Zuiker, 2003). This "video feedback" method initially seemed quite helpful and the process of selecting the clips helped the team appreciate important aspects of feedback conversations. The individual turns in the recorded feedback conversations were subsequently coded as being *off-task*, *neutral*, *procedural*, *factual*, *argumentation within GenScope*, and *argumentation beyond GenScope*. Analyses revealed only modest improvements in discourse (i.e., greater proportion of argumentation) in the classes that received video feedback, and students in those classes did not show larger gains on the NewWorm assessment. While the authors came away convinced that video-based scaffolding had value, the modest gains along with concerns over privacy and logistics in using actual video convinced us to pursue other strategies.

Of course, there are a vast range of tools and techniques that educators and researchers have advanced for supporting classroom discourse. The lessons from the initial video feedback study have been used to create animated video-coaches for specific close-level assessments that illustrate high- and low-quality enactments of that specific feedback conversation (Taasobshirazi, Zuiker, Anderson, & Hickey, 2006). We have experimented with a wide range of tools, following ideas

from previous research, prior implementations, and the teachers. The central point of this study is that the context of the close-level assessments, the opportunity to refine and test strategies across classes and/or assessments, and the evidence provided by proximal-level performance provide an ideal context for using and systematically refining these strategies.

We have also tried out a range of “conversation rubrics” and associated activities, both with and without video-based examples. In the elementary mathematics project, we are currently refining a rubric that defines the four aspects of group discourse (*explaining, listening, challenging, and reflecting*). After each feedback conversation, students review the rubric, informally reflect on their group’s discourse along each dimension, and select an aspect of their conversation to work on the next time. Initial results suggest that the rubric is helpful in supporting the students’ “reflexive awareness” about their discourse and has promise as a pedagogical tool for promoting and understanding discursive practices (Anderson, 2007).

Even as refinements become more focused on enhancing individual understanding in the subsequent cycles (discussed in subsequent sections), it is important to continue attending to the collective aspects of student learning. For example, the act of selecting clips for the video feedback study described previously helped us realize that the ways students negotiate transitions between items during feedback conversations crucially affected the overall quality of the discourse. Not surprisingly, groups tended to move on once they had reached consensus on the correct answer. This led to a core strategy of encouraging groups to stay with an item until every member had convinced the group that they understood everyone else’s reasoning and, hopefully, why some of the answers were more accurate than others. The specific point in this example is that the manner in which students collectively negotiated the routine of the feedback conversation also represented a form of learning, and a focus on reasoning over answers-as-products supported this learning. Had we focused prematurely on individual concept learning during the feedback conversations, this collective aspect of learning that may ultimately be more important might have been overlooked. We will return to this point in the conclusion.

One crucial decision is whether to pursue distal-level achievement data during the initial implementation cycle. In our studies, we have continued to do so in moderation by first constructing distal-level proxy tests from assembled pools of released items that are aligned to targeted standards and then randomly sampled to construct tests that can be used

in pre-post designs. These tests are used to ensure distal-level impact prior to evaluating that impact on standardized criterion-referenced tests. Constructing the tests forces the research team to grapple with the content standards that the curriculum will target and how they are manifested on other distal-level tests. However, just as the summative functions of external tests undermine the formative potential of classroom assessments (Black & Wiliam, 1998), focusing on distal-level evidence can undermine the efforts to scaffold collective discourse (by prematurely focusing on individual understanding). Furthermore, gains on distal assessments (and possible proximal assessments as well) are likely to be disappointing in the initial implementation cycle.

In the second GenScope assessment project, the students in the focal teacher's four classrooms during the initial implementation cycle gained 0.65 SD on the proximal-level NewWorm assessments. This was substantially less than the gains routinely obtained at the end of the first project, but more than double the 0.25 SD gains on the NewWorm later documented in the comparison classrooms at the same school. However, the GenScope teacher's students gained just 0.21 SD on the distal-level achievement test, less than half the 0.57 SD distal gain in the comparison classrooms. Similar findings were obtained in the initial implementation cycles across the NASA projects and with the Quest Atlantis project (Barab, Sadler, Heiselt, Hickey, & Zuiker, 2007).

The goals of the implementation cycle illustrate how a discursive approach to assessment provides a potentially useful extension to conventional notions of accountability. The goal of this cycle is to leave students and teachers with useable knowledge about how to participate in productive discourse during the curricular activities and close-level assessments, helping them to use that knowledge to continually improve that discourse. Specifically, classrooms should appreciate that individuals will excel on proximal assessments (e.g., formal exams) if they first work together to support each other's participation in discourse and argumentation around close-level assessments and activities. It is in this sense that this first cycle should establish informal "student-oriented" accountability where students hold each other accountable for their collective participation in domain-specific classroom discourse.

Experimentation cycle. The next refinement cycle defines a scalable suite of activities, assessments, and scaffolds that ultimately has resulted in the largest gain on students' distal-level achievement. The various tools and strategies that emerged in the implementation cycle are more formally examined to confirm their value and support further refine-

ment. Ideally, additional teachers are asked to implement, using a more sustainable level of research resources and support. Working with implementation teachers, we design appropriately complex studies to help define a version of the approach that is both scalable and has the largest impact on achievement.

One pressing issue that we are currently exploring in the elementary mathematics project concerns feedback on the proximal-level assessments. As curriculum-oriented assessments, they provide teachers with useful evidence about certain ways each student understands the core concepts and skills targeted in the curriculum. They also provide the research team with useful evidence about the impact of the activities and close-level assessments. In addition to supporting more formal remediation by the teachers, proximal assessments also have formative potential for confirming or adjusting students' understanding or misunderstanding of targeted concepts. Specifically, we presume that feedback conversations following students' completion of the proximal assessments might ensure that the students' learning of concepts in curricular contexts transfers to a range of subsequent contexts, including high-stakes tests, because of the deepened understandings we think such discourse and reflection provide. For example, careful assembly of proximal-level items and careful wording of the corresponding answer explanations can help students appreciate common misconceptions and see how item writers exploit those misconceptions to prompt some students to select the incorrect response.

Despite the potential of proximal-level discursive feedback, it turns out to be a challenging activity to support, and it may confound other goals for the proximal-level evidence. Therefore, we are attempting to gather evidence of its distal-level impact. In the elementary mathematics project we have attempted several quasi-experimental designs, including within-class/between-student designs that counterbalance the order of the proximal assessment and its feedback with the distal test, and examine whether test scores are higher when they follow proximal-level feedback. An alternative design that requires fewer students but more testing time is administering the test before *and* after the proximal assessment and feedback. We have already documented modest gains on distal tests with such designs (Hickey & Cross, 2006). Given that distal-level assessments by their nature are quite insensitive to any short-term interventions and that the proximal feedback conversations only lasted 30 minutes, we are now vigorously pursuing this aspect of our innovation in other current projects.

Another feature of the experimentation cycle—adding additional implementation teachers—naturally leads the research team to more formally address professional development as well. The feedback conversations provide a useful context in which the enormous range of prior research on professional development can be considered. Reflecting our focus on classroom discourse, we have found that resources for language arts education (e.g., Leung & Mohan, 2004) are particularly relevant. Video technology is promising for professional development as well because it can provide teachers with salient examples and benchmarks of ideal (and problematic) enactments of feedback conversations. Building on the inspiration and ideas of others (e.g., Sherin & Han, 2004), we have experimented with quite a few strategies in this regard. One key finding alluded to earlier is that “live” video collected during a research project presents significant issues of privacy and consent. While negative examples are powerful tools for helping educators appreciate the nuances in the positive examples, unflattering clips from actual classrooms cannot be incorporated into professional development. A promising alternative has been to work with research participants to “reenact” feedback conversation. Such video is collected under the auspices of a dramatic event. As reenactments make modest demands on acting ability, they may be a promising alternative to the prohibitively expensive dramatic productions (Hickey, Wallace, Hay, & Recesso, 2002).

In terms of accountability, the implementation cycle aims to establish “student-centered” accountability around the close-level assessments. In contrast, the experimentation cycle should establish a “teacher-centered” accountability around the proximal-level assessments. This teacher-centered accountability is “overlaid” on top of the student-centered accountability; in other words, while the students are responsible for excelling on the close-level assessments and feedback conversations, the teacher is responsible for ensuring that students excel on the proximal-level assessment and providing corresponding remediation for specific students and/or topics. Ultimately, in light of the iterative cycles of refinement, it is the research team’s responsibility to ensure that students in classrooms where both forms of accountability are established also excel on the distal-level achievement test. By the second year of the second GenScope Assessment Project, the focal GenScope teacher had attained gains of 1.5 SD on proximal-level New-Worm (as large as those found in most GenScope classes in the first project). Most importantly, these same students gained 0.74 on the distal-level achievement tests, which was larger than the 0.57 distal-

level gains in the comparison classes. In the NASA project, similar gains were obtained for two of the three curricular packages after two cycles (the extent of the implementation). Such findings warrant larger-scale implementation and more formal evaluation in a subsequent evaluation cycle.

Evaluation cycle. The final cycle considers the entire suite of activities, assessments, and scaffolds in terms of external achievement measures. However, the scope of the evaluation depends on the scope of the project and the resources available. Again, the presence of multiple levels of assessments and a design-based approach offers numerous possibilities for obtaining useful evidence. In most cases, a comparison group is appropriate. The aforementioned comparison teacher in the second GenScope study had completed more university coursework in genetics and had more years of teaching experience than the focal teacher. Additionally, he devoted the same number of class periods to introductory genetics and used the same district-mandated textbook that the GenScope teacher would have been using.

In the final cycle of that project, the students in the four focal classes gained an average of 2.0 SD on the proximal level NewWorm assessment. Providing the most convincing evidence obtained so far, these same students gained 1.1 SD on the distal-level achievement test, which was about double the distal-level gains in the two matched comparison classrooms. However, we believe that it is even more convincing that the pattern of increasingly large annual gains on the distal-level test clearly “echoes” (i.e., mirrors, but to a smaller extent) the annual increases on the more directly targeted proximal-level NewWorm assessment.

Our elementary mathematics project will provide the most rigorous evaluation of this research cycle so far. The project is developing close-level and proximal-level classroom assessments for the entire year’s mathematics curriculum. It is being scaled up to include all fifth-grade teachers in two implementation schools and gains will be formally evaluated against all fifth graders in two closely matched comparison schools. A comprehensive evaluation will provide evidence about three additional consequences of raising distal-level scores in this fashion. First, the study will examine whether or not distal-level gains in achievement are echoed (i.e., transfer) to remote-level achievement by examining student performance on corresponding subtests on the norm-referenced achievement test. Second, the study will examine distal-level gains in individual understanding and collective discourse by

having every student complete carefully selected performance assessments (i.e., aligned to the standards and not the curriculum) and by conducting discourse analysis of representative triads of students collaboratively solving similar problems. Finally, the project will assess the broader usefulness of the entire set of materials and practices by having all of the comparison teachers implement them after the comparison data have been collected and by comparing gains from one year to the next on the entire set of outcomes.

Conclusions

Our ultimate goal is to create a self-sustaining framework featuring high-quality assessments and activities that are aligned coherently with external tests for use by teachers who have acquired the skills needed to develop and refine their own close-level assessments. With such assessments in place, students, teachers, and administrators should then be able to work together in continual “evidence-based” educational practice that actually delivers meaningful improvements while adapting to inevitably changing educational goals. Of course, doing this will require broader changes in school culture, teacher professional development, and classroom and external accountability practices. We believe the approach that we have outlined here offers a useful trajectory, consistent with both contemporary assumptions about worthwhile classroom instruction *and* current accountability-oriented school reforms, for doing so.

To reiterate, our approach is shaped by sociocultural perspectives that view all learning as social change. It is acknowledged that our socioculturally oriented approach presents theoretical challenges for many readers interested in using evidence to improve education. Arguably, including distal-level outcomes and conventional evaluation methods in our research trajectory addresses the tensions between sociocultural views and conventional individually oriented views of learning. It does so by transforming essentially philosophical tensions into practical questions that can be solved empirically using widely appreciated methods. Ultimately, however, we have concluded that an appreciation of sociocultural views of learning is necessary to fully appreciate and exploit the value of such approaches. For example, we previously suggested that discursive assessment practices could be undermined by prematurely focusing on the learning of individual students. This characterization actually underrepresents the concern. In a very important way, we *never* truly focus on “individual” learn-

ing. When all learning is viewed as social change, the act of completing any assessment is viewed as participation in collective discourse—albeit a specific form of discourse (Gee, 2003; Hickey & Zuiker, 2003). Therefore, increased test scores are viewed as evidence of increasingly successful participation in what we ultimately understand to be fundamentally *social activity*.

Other chapters in this volume have discussed the many challenges facing evidence-based educational reform. We close by considering two of these challenges and suggest that sociocultural perspectives and a multilevel framework have unique potential for understanding and addressing them. The first challenge concerns the controversies over competing assessment formats, from multiple choice to more open-ended formats to group level forms of assessment, and the conceptions of learning that underlie them. We contend that a sociocultural perspective assumes that the act of completing any type of assessment is a “special case” of socially situated activity. In other words, different assessments support specialized forms of discourse, which are necessary to provide different forms of evidence that have different utility. Furthermore, we contend that such a perspective, along with the differentiated view of ideal functions of different assessment levels, may help clarify when various item formats are more or less useful.

A second challenge to using evidence to improve education concerns the validity of gains on targeted tests as evidence of broader educational improvement. One of the main concerns with the No Child Left Behind Act is the evidence that increased scores on targeted criterion-referenced tests are often associated with declining scores on other nontargeted tests, such as college placement tests and the NAEP (e.g., Ghezzi, 2006; Winerip, 2005). This evidence is stoking concerns that excessive pressure to directly raise test scores will lead to a narrowing of the curriculum and diminished coverage of topics or types of understanding not included in the targeted test (e.g., Burroughs, Groce, & Webeck, 2005). Our multilevel approach reflects our belief that efforts to increase performance on targeted assessments should be associated with corresponding (but smaller) increases at a subsequent, more distal-level of outcomes. We strongly believe that the educational value of competing evidence-based educational reforms should ultimately be evaluated by considering their impact on more distal, nontargeted outcomes.

We are very encouraged by initial evidence that brief discursive feedback on proximal assessments supports distal-level gains and believe that much of the instructional time and money now being devoted to test-prep training programs could be usefully redirected in this manner.

This could be quite readily accomplished with some of our existing assessments within existing NCLB-mandated after-school tutoring programs. The setting lends itself well to random assignment, which we expect would provide rigorous evidence about the limited impact of test-prep programs and the broader advantages of a more discursive approach. Of course the ultimate goals are more ambitious, and will require much broader consideration and debate. Focused efforts like the ones summarized in this chapter are important first steps, which we hope readers will find thought-provoking and worthy of further consideration.

AUTHORS' NOTE

The primary studies described in this chapter were supported by grants RED-955348 and REC-0196225 from the National Science Foundation. The opinions expressed here are those of the authors and do not necessarily reflect the opinions of the National Science Foundation. We wish to acknowledge the contributions of the individuals listed on referenced publications for their contributions to the work described here, as well as the input of teachers and students who participated in this research.

NOTE

1. The GenScope software and all of the assessments described here are available from the Concord Consortium at <http://genscope.concord.org/research/>.

REFERENCES

- Anderson, K.T. (2007, April). *Discursive meta-tools for the development of practice and identity in an elementary math classroom*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Anderson, K.T., Zuiker, S., Taasoobshirazi, G., & Hickey, D.T. (in press). Discourse analysis for enhancing the formative value of classroom assessment practices in science. *International Journal of Science Education*.
- Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *The Journal of the Learning Sciences, 13*, 1–14.
- 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., Sadler, T., Heiselt, C., Hickey, D., & Zuiker, S. (2007). Relating narrative, inquiry, and inscriptions: Supporting consequential play. *Journal of Science Education and Technology, 16*, 59–82.
- Beach, K. (2003). Learning in complex social situations meets information processing and mental representation: Some consequences for educational assessment. *Measurement, 1*, 149–177.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*, 7–74.
- Burkhardt, H., & Schoenfeld, A.H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher, 32*(9), 3–14.
- Burroughs, S., Groce, E., & Webeck, M.L. (2005). Social studies education in the age of testing and accountability. *Educational Measurement: Issues and Practice, 24*, 13–20.

- Case, R. (1996). Changing views of knowledge and the impact on educational research and practice. In D.R. Olson & N. Torrance (Eds.), *The handbook of education and human development* (pp. 75–99). Malden, MA: Blackwell Publishers.
- Cobb, P., Confrey, J., DiSessa, A., Lehrer, R., & Schauble, L. (2003). Design experiments in educational research. *Educational Researcher*, 32(1), 9–13.
- Duschl, R.A., & Gitomer, D.H. (1997). Strategies and challenges to changing the focus of assessment and instruction in science classrooms. *Educational Assessment*, 4(1), 37–73.
- Gee, J.P. (2001). Educational linguistics. In M. Aronoff & J.R. Miller (Eds.), *Handbook of linguistics* (pp. 647–663). Malden, MA: Blackwell Publishing.
- Gee, J.P. (2003). Opportunity to learn: A language-based perspective on assessment. *Assessment in Education*, 10, 25–44.
- Ghezzi, P. (2006, August 31). Report: Georgia student tests are too easy. State works to revise standards for achievement. *Atlanta Journal Constitution*, p. 1.
- Greeno, J.G., & the Middle School Mathematics through Application Project Group (MMAP). (1998). The situativity of knowing, learning, & research. *American Psychologist*, 53(1), 5–26.
- 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.
- Haertel, E.H., & Greeno, J.G. (2003). A situative perspective: Broadening the foundations of assessment. *Measurement*, 1(2), 154–162.
- Hickey, D.T. (2001). *Assessment, motivation, & epistemological reconciliation in a technology-supported learning environment*. Grant REC-0196225 from the National Science Foundation to the University of Georgia.
- Hickey, D.T. (2003). *Design-based implementation and evaluation of NASA CET multimedia science curriculum*. Subcontract from the Wheeling Jesuit University Center for Educational Technology to the University of Georgia.
- Hickey, D.T., & Cross, D.I. (2006, April). *Design-based multi-level assessment for enhancing discourse, learning, curriculum, and achievement in elementary mathematics*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.
- 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., & Zuiker, S. (2003). A new perspective for evaluating innovative science learning environments. *Science Education*, 87, 539–563.
- 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, 495–538.
- Hickey, D.T., Mewborn, D.S., & Lewison, M.A. (2005). *Multi-level assessment for enhancing mathematical discourse, curriculum, and achievement in diverse elementary school classrooms*. Grant REC 0553072 from the U.S. National Science Foundation to Indiana University.
- Hickey, D.T., Wallace, C., Hay, K., & Recesso, A. (2002). *Video-supported formative assessment of inquiry-oriented activity and instruction*. Grant from the University of Georgia Professional Preparation of Educators Mini-Grant Program to the UGA Learning and Performance Support Laboratory.
- 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, 155–196.
- Hickey, D.T., Zuiker, S.J., Taasoobshirazi, G., Schafer, N.J., & Michael, M.A. (2006). Three is the magic number: A design-based framework for balancing formative

- and summative functions of assessment. *Studies in Educational Evaluation*, 32, 180–201.
- 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: Lawrence Erlbaum Associates.
- Jiménez-Aleixandre, M.P., Rodríguez, A.B., & Duschl, R.A. (2000). “Doing the lesson” or “doing science”: Argument in high school genetics. *Science Education*, 84, 757–792.
- Lemke, J.J. (2000). Across the scale of time: Artifacts, activities, and meaning in ecosocial systems. *Mind, Culture, and Activity*, 7, 273–290.
- Leung, C., & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: Assessment as discourse and assessment of discourse. *Language Testing*, 21, 335–359.
- O’Connor, M.C. (2001). “Can any fraction be turned into a decimal?” A case study of a mathematical group discussion. *Educational Studies in Mathematics*, 46, 143–185.
- 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.
- Schafer, N.J., Kruger, A., Hickey, D.T., & Zuiker, S. (2003, April). *Using video feedback to facilitate classroom assessment conversation*. Paper presented at the annual meeting of the American Educational Research Association, Chicago.
- Sherin, M.G., & Han, S.Y. (2004). Teacher learning in the context of a video club. *Teaching and Teacher Education*, 20, 163–183.
- Taasobshirazi, G., Zuiker, S.J., Anderson, K.T., & Hickey, D.T. (2006). Enhancing inquiry, understanding, and achievement in an astronomy multimedia learning environment. *Journal of Science Education and Technology*, 15, 383–395.
- Vygotsky, L.S. (1986). *Thought and language*. Cambridge, MA: MIT Press.
- Wenger, E. (1998). *Communities of practice: Learning, meaning, & identity*. Cambridge: Cambridge University Press.
- Wiggins, G.P., & McTigue, J. (2005). *Understanding by design*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Winerip, M. (2005, November 2). Are schools passing or failing? Now there’s a third choice ... both. *New York Times*, p. 1.
- Zuiker, S.J., Hickey, D.T., Kwon, E.J., Chapman, R., & Barab, S.A. (2005, August). *Assessing student learning in, around, and for a multi-user virtual environment*. Presentation at the bi-annual conference of the European Association for Research on Learning and Instruction, Nicosia, Cyprus.