CHAPTER 7

# THEORY, LEVEL, AND FUNCTION

## Three Dimensions for Understanding Transfer and Student Assessment

**Daniel T. Hickey**
***University of Georgia***

**James W. Pellegrino**
***University of Illinois, Chicago***

**ABSTRACT**

Assessment is central to education and the study of learning. Most educational assessments measure the transfer of learning. As such, the study of transfer is broadly relevant to educational assessment, and vice versa. This chapter uses transfer as a vehicle for understanding and improving multiple forms of student assessment including classroom assessments used by teachers and external assessments such as standardized achievement tests. This also includes assessments used to measure the knowledge of individuals and to measure the effectiveness of particular learning environments. Our consideration of assessment and transfer uses a framework that consists of three interrelated dimen-

*Transfer of Learning from a Modern Multidisciplinary Perspective*, pages 251–293

sions. The first dimension concerns *theory* of knowing and learning (*empiricist, rationalist,* or *socioculturalist*). The second dimension concerns *level* of assessment relative to a particular learning environment (*immediate, close, proximal, distal,* or *remote*). The third dimension concerns *function* of assessment practice (*formative* and *summative*). We extend prior consideration of these three aspects of assessment by framing each in terms of knowledge transfer and then considering their complex interactions. Examples of the dimensions and their interaction are taken from typical assessment practices, the research literature, and a program of research carried out in part to help define the dimensions. We explore the unique formative and summative functions at each assessment level and show how different theories are needed at different levels to balance assessment functions within and across levels. We argue that doing so facilitates the crucial alignment of assessment practices across levels. This should help accomplish the diverse goals of educational assessment and address major tensions associated with accountability-oriented educational reforms as well as stringent new criteria regarding sources of evidence in educational research.

## INTRODUCTION

*Transfer of learning*, or simply *transfer*, refers to the use of knowledge learned in one context in some other context. The chapters in this volume illustrate the continuing interest in transfer among researchers concerned with learning and education. As will be argued, issues of transfer are particularly relevant to understanding the practices of educational *assessment*. The converse is also true since assessment always constitutes a test of transfer. This chapter uses transfer as a vehicle for understanding and improving the many forms of student assessment that are common in education. This includes *classroom* assessments, such as the quizzes and exams used by teachers, as well as *external* assessments, such as standardized achievement tests used for high-stakes accountability purposes. The chapter considers transfer in the context of assessments used to measure the knowledge of individuals, as well as assessments used to measure the effectiveness of particular learning environments. We are motivated by the desire to establish a comprehensive framework for understanding and resolving the many issues in educational research and practice that involve assessment and that therefore implicate issues of transfer.

Our consideration of the relationship between transfer and assessment uses a framework that consists of three interrelated dimensions: *theory* of knowing and learning, *level* of relationship to a particular learning environment, and *function* of assessment practice. Each of these dimensions has been discussed in the assessment literature and may be familiar to many readers. This chapter extends prior considerations of each of these dimensions by framing each in terms of knowledge transfer issues, and then con-

sidering the complex interactions among the various dimensions. The chapter also attempts to illustrate these dimensions and their interactions using general types of assessments that are familiar to most readers, as well as specific examples drawn from the research literature. Additional examples are drawn from an assessment research project (Hickey, 2001) involving the *GenScope* computer-based curriculum for introductory genetics (Horwitz & Christie, 2000). These examples are particularly relevant because the project was designed in part to help specify the same three dimensions that are detailed in this chapter.

The choice of dimensions and the nature of each are intended to be agnostic relative to prevailing practices of teaching and assessment. While our own biases will be apparent in the discussions and examples, our concern is with highlighting how systematic consideration of the three dimensions can help address pressing policy and practice challenges in the current K–12 educational context. For example, the chapter considers the meaning of recent accountability-oriented reforms in some detail in light of the transfer–assessment relationship, but does not take any particular position on the appropriateness of such practices.

The next section elaborates the chapter's goals and is followed by a detailed consideration of each of the three dimensions and some examples of their interactions. The chapter concludes with a consideration of additional research and interpretation issues that demand attention if we are to improve education via a clearer understanding of the complex issues surrounding the design and use of educational assessments in both educational practice and research. To help orient the reader, Table 7.1 presents the three dimensions in our framework and the contents of each.

**Table 7.1. Three Dimensions of Transfer for Assessment**

| *Assessment of Transfer Dimension* | *Contents of Dimension* | *Description/Example of Dimension Contents* |
|---|---|---|
| Theories of knowing and learning (i.e., assumptions about transfer) | *Empiricist* | Transfer of associations |
| | *Rationalist* | Transfer of schema |
| | *Socioculturalist* | Transfer of participation |
| Levels of assessment (i.e., distance from a particular learning environment) | *Immediate* | Informal observations |
| | *Close* | Semi-formal classroom assessment |
| | *Proximal* | Formal classroom assessment |
| | *Distal* | Criterion-referenced tests |
| | *Remote* | Norm-referenced tests |
| Functions of assessment (i.e., educational goal of a particular assessment practice) | *Formative* | Advancing knowledge and learning |
| | *Summative* | Measuring knowledge and learning |

## CHAPTER OVERVIEW AND GOALS

The framework outlined in this chapter is designed to highlight assumptions about knowledge transfer that are often taken for granted when assessing students' knowledge. As pointed out by Barnett and Ceci (Chapter 8, this volume), education is fundamentally about preparing learners for life beyond schooling. Assessment is ubiquitous in education because it is presumed to provide useful evidence of success in this regard. Therefore, most student assessment ultimately serves as a proxy to predict whether school learning will transfer to some other "real-world" context (further schooling, work, citizenship, etc.). Stated in the broadest possible terms, assessment assumes that the knowledge that affords successful participation in the assessment context will also afford successful participation in some other context. As such, our assumptions about the nature of knowledge are important to understanding and improving the assessment of knowledge transfer. Thus, our chapter's first core goal is *uncovering how different views of knowing and learning lead to different assumptions about transfer, and detailing the implications of these assumptions for assessment.*

We pursue this first goal in part by building on other similarly inspired efforts. For example, an expert panel at the U.S. National Research Council (NRC) released a report in 2001 entitled *Knowing What Students Know: The Science and Design of Educational Assessment.* The report characterized assessment in the following manner:

> Every assessment, regardless of its purpose, rests on three pillars: a model of how students represent knowledge and develop competence in the subject domain, tasks or situations that allow one to observe students' performance, and an interpretation method for drawing inferences from the performance evidence thus obtained (2001b, p. 2)

The report points out that that each pillar of the "assessment triangle" is necessarily based on theoretical assumptions, but argues that these assumptions are often implicit and taken for granted. Perhaps more seriously, and particularly in the case of large-scale, standardized assessments of student academic achievement, the report argues that these assumptions are "based on highly restrictive beliefs about learning and competence not fully in keeping with current knowledge about human cognition and learning" (NRC, 2001b, p. 2). In this regard, the report provides further support for the long-standing argument that (1) many assessment practices are based on outdated and/or highly restricted theoretical assumptions, (2) that these assumptions need to be acknowledged and challenged, and (3) that newer theories of knowing and learning have untapped potential for improving the quality and use of assessment in

educational settings (e.g., Gipps, 1994; Mislevy, 1993; Pellegrino, Baxter, & Glaser, 1999; Resnick & Resnick, 1992). Thus, one part of our first core goal is considering the broader range of theories of learning and transfer that might underlie assessment practices as part of explicating the first dimension in the framework.

## Applying a Broader View of Knowing and Learning

Situative and sociocultural views of knowing and learning challenge 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 in and fundamentally residing in ritualized cultural practices. *Knowing What Students Know* (NRC, 2001b) and other recent calls to broaden assessment (e.g., Mislevy, Stienberg, & Almond, 2002) certainly acknowledge these views. Arguably, though, their consideration and recommendations for assessment still take for granted core assumptions about knowing and learning that some consider too narrow. For example, discussion of the development of knowledge in *Knowing What Students Know* is largely based on modern cognitive notions such as schema theory, expert–novice differences, and conceptual change. Notions of context and language that have unique significance in the sociocultural view are less prominent or considered within a modern cognitive/rationalist perspective. Transfer is characterized primarily in terms of the knowledge that individuals develop in learning situations that enables them to succeed in subsequent transfer situations. Perhaps it is not surprising that sociocultural views have as yet had minimal impact on discussions of assessment practice, given that such views typically reject the individually oriented view of knowledge that underlies nearly all prior conceptualization of assessment (e.g., Delandshere, 2002; Gipps, 1999).

That does not mean that sociocultural views do not offer ways of addressing the broader issues that assessment has traditionally targeted. However, the articulation of how one does so has been lacking among proponents of this theoretical persuasion. Thus, an essential aspect of this first core goal is *summarizing uniquely sociocultural assumptions about transfer, and identifying their implications for assessment.* We do so by drawing on an approach used in handbook chapters by Greeno, Collins, and Resnick (1996) and Case (1996) that compares three "grand theories" of knowing and learning, using the widely accepted labels of *empiricist, rationalist,* and *socioculturalist.* In order to identify the unique views of assessment and transfer associated with each, we start with the core assumptions about knowing and learning within each view. This reveals core assumptions

about learning and transfer that in turn make it possible to identify assessment practices that are more and less consistent with each perspective.

## Defining Multiple Levels of the Assessment–Transfer Relationship

Many assessment practices, including most large-scale, standardized achievement tests, are designed to be independent of any particular curriculum or learning environment. In such cases, the transfer assumption is that the assessed knowledge exists in a form that can be tapped by the test—that is, will readily transfer to use in the testing context—and will indeed then transfer to use in some subsequent real-world environment. Such a simplistic notion of transfer and the connections among learning contexts, assessment contexts, and contexts beyond the classroom has led to many arguments about the appropriateness of certain forms of assessment as valid indices of the effects of instructional programs and innovations such as inquiry-based learning environments, many of which involve various technology-based materials and tools (see, e.g., Pellegrino, 2004).

As we will illustrate, multiple transfer assumptions must of necessity come into play when assessing students' knowledge relative to particular learning environments. This is usually the case when researching or evaluating innovative curriculum or instructional strategies. Stated broadly, such assessment requires the additional assumption that new knowledge that results from successful participation in the learning environment will lead to more successful participation in the assessment context. These additional assumptions make the assessment of knowledge transfer relative to a specific learning environment quite complex. As such, the typical distinction between *zero, near,* and *far* transfer does not provide sufficient precision. Thus, our chapter's second core goal is *defining the relationship between transfer and assessment level, in terms of the "distance" of a particular assessment from a particular educational context, and considering how assessment level interacts with other dimensions of assessment.*

Most readers will be familiar with the distinction between two levels: "internal" classroom assessment and "external" testing. More recently, assessment theorists have defined a larger number of levels. We start with the five levels defined in the consideration of summative assessment functions by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002): *immediate, close, proximal, distal,* and *remote.* We then extend this understanding by considering its relationship with the other two dimensions. Doing so highlights the nuanced distances *within* levels, in terms of the distance between a particular assessment and feedback regarding performance on that assessment. It turns out that altering the format of the feedback relative to the

format of the assessment (i.e., "increasing the distance") can increase the formative value of that feedback for improving the transfer of learning and instruction.

In general, we contend that the notion of levels is directly relevant to broader considerations of knowledge transfer, because efforts to measure knowledge transfer almost always involve assessment. We further contend that the notion of levels is central to accomplishing our first goal. This is because of the seemingly fundamental relationship between levels of assessment and theories of knowing and learning. For example, we show that goals of assessment practices at the *immediate level* are well suited to newer sociocultural views of knowing and learning, but ill-suited to traditional behavioral views.

## Balancing the Varied Functions of Assessment

One of the central points of *Knowing What Students Know* was that assessments are developed for specific purposes and the nature of their design is very much constrained by their intended use. This point about the reciprocal relationships between function and design leads to concerns about the inappropriate and ineffective use of assessments for purposes beyond the original intent. This concern is most obvious in the context of the distinction between classroom assessments and large-scale external tests. Assessments that serve the more summative functions of external assessments are usually inappropriate for the more formative functions of classroom assessment. The converse holds as well. Thus, our chapter's third core goal is *providing a more detailed understanding of assessment functions that goes beyond familiar distinctions between formative and summative functions and between classroom and external assessments, and considering how those functions interact with other dimensions of assessment.* We accomplish this by defining the whole range of summative and formative functions for every level of assessment practice, and considering whether different theories of transfer are more appropriate for particular levels.

Our broader consideration of assessment function is intended to advance thinking about the formative functions of assessment. Formative classroom assessment was considered in another National Research Council report (NRC, 2001a) that endorsed the educational potential of such assessments, while voicing several areas of concern. For example, it was pointed out that different forms of assessment and different types of assessment items vary in their capacity to support a formative assessment goal, and that much of the potentially useful information from classroom assessment is unused, or used in ways that undermine learning and instruction.

The framework we advance is also relevant to understanding variation in the likelihood of success of efforts to use external large-scale testing to improve teaching, learning, and achievement. This includes the large-scale testing reforms such as performance assessments that many states in the United States implemented during the 1990s (e.g., Herman, 1997). It also includes more recent accountability oriented reforms that are transforming the educational landscape in many countries, including the United States. For example, the No Child Left Behind Act of 2001 (NCLB) requires schools to continually increase scores on criterion-referenced achievement tests administered by each state. This legislation demands such increases for all students, including those who are economically disadvantaged; from racial, ethnic, and linguistic minorities; and those who have been identified as suffering from behavioral or learning disabilities. Furthermore, NCLB demands that these gains not come at the expense of other desirable educational outcomes, such as performance on other criterion-referenced tests (e.g., advanced placement tests), norm-referenced tests (e.g., Iowa Test of Basic Skills & Scholastic Achievement Test), graduation rates, college success, and so on. To varying degrees, stakeholders worry about other outcomes that are not as easily measured, such as deep conceptual knowledge, graduation rates, advanced course offerings, initial college success, and so on. Many of the controversies about accountability ultimately concern the validity of various assessments as evidence that individuals have knowledge that will actually transfer to subsequent settings (Messick, 1995)

To meet our third goal, the chapter starts from the assumption that all assessments, ranging from informal internal classroom assessments to external, standardized achievement tests, are designed for particular purposes, and must be understood in the context of their intended use *and* actual use (NRC, 2001b). The chapter then systematically considers the formative and summative potential of assessment practices at all five levels. The chapter closes by arguing that systematically shifting the way that knowledge is represented across multiple assessment levels (i.e., from socioculturalist to rationalist to empiricist) is a promising means of increasing the value of assessment for meeting broad and potentially conflicting goals for education.

## COMPETING THEORIES OF KNOWING AND LEARNING, TRANSFER, AND ASSESSMENT

Assessment theorists have long argued against the misconception that narrowly equates "knowledge" and (therefore) "learning" with the specific knowledge elements typically assessed with multiple-choice achievement

tests. Ruiz-Primo and Shavelson (1996) distinguished between *declarative, procedural,* and *strategic* knowledge in their consideration of different forms of assessment. A central conclusion of *Knowing What Students Know* (NRC, 2001b) was the need for a broadened view of assessment that more completely embraces modern insights about knowing and learning from cognitive science. Following from analyses by Greeno and colleagues (1996) and Case (1996), our framework's first dimension supports an even broader view. We consider the unique views of assessment and transfer that follow from the three "grand theories" of knowing and learning, which we choose to label *empiricist, rationalist,* and *socioculturalist.*

This first dimension should not be viewed as an effort to advance one theory of transfer over another, or as a summary of the current state of understanding of transfer within each theory. Rather, this dimension is advanced as a useful way of understanding different views of transfer in order to better explicate and address practical issues in assessment, testing, and evaluation that follow from those differences. We acknowledge that some readers may find the following characterizations overly narrow, bordering on caricature. This is most likely to be the case with readers who envision broader models of assessment practice without a corresponding broadening of their views of knowing and learning. Zimmerman (1993) has argued that such "comparative" analyses overemphasize initial conceptualizations of competing theories and ignore their evolution. It is possible, however, that such concerns are most relevant when one's goal is refining and validating broad psychological theories within conventional basic research paradigms. While the comparative analysis itself springs from an unabashedly partisan goal of identifying the unique affordances of newer sociocultural theories for assessment practice, the motivation for such an approach is the desire to refine effective and inspiring models of educational practice (e.g., Collins, 1999; Lagemann, 1999). In this regard, the value of this first dimension in our framework is premised on two arguments. The first is that tensions between empiricist and rationalist perspectives, as detailed below, are a primary obstacle in the advancement of assessment practice. The second argument is that it is necessary to identify the unique and defining characteristics of all three perspectives in order to consider the suggestion that sociocultural perspectives offer a unique means of overcoming seemingly intractable tensions between the first two perspectives.

## An Empiricist Perspective on Assessment of Transferable Knowledge

Empiricist views of knowing and learning are embodied in the behaviorist models associated with Skinner and later in the human information-

processing models following the "cognitive revolution" in the 1960s. Empiricist views of teaching are epitomized in traditional teacher-directed models of teaching, and are most consistent with naive "folk psychology" perspectives on learning (Bereiter & Scardamalia, 1996). Empiricist perspectives dominated formal models of schooling for the first half of the 20th century. While their influence in education waned with the emergence of rationalist perspectives in the 1970s, they are reemerging within the context of "back-to-basics" movements and competitive market-based educational reforms, and the associated resurgence of conventional direct-instruction teaching practices.

***Knowing as Having Associations***

Rooted in the British empiricist philosophy of Hume, empiricist perspectives view the mind as a device for detecting and operating on patterns in the world. The "world" is construed as an objective knowable reality. This means that there is some ultimate "truth" to be known. In the earlier behavioral empiricist models, the fragments that make up one's knowledge of a domain are construed as stimulus–response associations; in subsequent cognitive-associationist empiricist models, such as Bloom's taxonomy (e.g., 1976) and Gagné's ISD models (e.g., 1985), the fragments that represent one's knowledge of a domain were construed as either *declarative* knowledge of facts and concepts (i.e., "patterns that can be detected") or *procedural* knowledge of skills and processes (i.e., "operations we can execute on those patterns"). Reflecting a mechanistic meta-theoretical perspective, empiricist perspectives are inherently *reductionist* (assuming that complex behavior or concepts consist of smaller elements) and *additive* (assuming these smaller elements then assemble into an accurate representation of the more complex entity). This means that the various "higher-level" elements are merely associations between lower-level elements, rather than some sort of more holistic entity. When individuals engage in some activity that is presumed to indicate that they know something (solving a problem, answering a question, etc.), it is because they appropriately marshal all of the needed lower-level components of knowledge and skill.

Relative to academic domains, an empiricist perspective views knowledge as knowable aspects of the natural world. Once scientists and scholars have discovered this aspect of the world, and agreed on its nature and value, it is the job of the schools to figure out which aspects of this knowledge are most important, and then impart that knowledge to youngsters. From this perspective, a student's knowledge of this domain of knowledge is an incomplete "imprinting" of whatever elements or aspects of the domain to which the learner has been exposed. This includes declarative knowledge and procedural knowledge. Each of these elements of knowledge is presumed to

consist of additional lower-level declarative and procedural knowledge. Traditional scope and sequence charts are one common representation of how such a perspective views the knowledge in this domain.

***Learning and Transfer as Acquiring and Applying Associations***

When knowledge is viewed as an organized collection of many small cognitive or behavioral associations, learning is seen as the process of forming, strengthening, and adjusting those associations. Learning occurs when we are exposed to patterns (i.e., associations) and become able to recognize and respond to those patterns efficiently. Learning continues as many smaller associations are assembled into larger ones. This suggests that ideal learning environments efficiently transmit the information represented by these associations, and then maintain student engagement in the routines needed to build and strengthen these associations. This process is best illustrated by the conventional "direct instruction" approach, where a discrete concept or skill is presented to students who then practice using that concept or performing that skill under conditions that motivate the student and optimize transfer. When formulating curricula in such an environment, the primary concern is sequencing from component to composite skills.

When knowing and learning are viewed in terms of having and acquiring associations, the transfer of learning to a new situation depends on the number and nature of the associations that are needed in the new situation, relative to the number and nature of the associations acquired in the previous environment. Because of the assumptions about reductionism and additivity, it is assumed that the components of knowledge transfer quite readily. By demonstrating that students have made associations that are generally agreed to be useful in some transfer environment, the students are presumed to have some transferable knowledge. Thus, demonstrating that students have mastered the various components of knowledge from a domain is presumed sufficient to ensure that they have transferable knowledge—presuming that those components are needed to solve problems in the transfer environment.

***Assessment as Testing of Components***

From this perspective, the assessment of transferable knowledge is carried out around the numerous specific associations that are assumed to comprise that knowledge. Because this perspective assumes that complex knowledge can and should be broken into its components, it is entirely appropriate to look for transfer of the components of knowledge that learners were exposed to in the learning environment. This can be done quite efficiently by asking students to recognize or recall those same associations, or to make new higher-order associations between existing lower-level associations. The familiar multiple-choice and short-answer formats

are well suited to assessing the transfer of knowledge in this regard. In classroom settings teachers routinely use such tests, often obtaining them from teacher versions of their textbooks or from supplementary materials such as test item banks.

Importantly, this is not arguing that assessing transfer with multiple-choice and short-answer formats requires one to embrace an empiricist view of learning and transfer. Rather, this is to say that formats that directly assess specific associations (or associations between specific associations) are the logical choice when assessing the transfer of learning from this perspective. The more general point here is that the practical choice of assessment format is not *dictated* by one's theory of knowing and learning. If one embraces an empiricist view of knowing and learning, the familiar multiple-choice formats that were refined in light of those views offer a sensible, if not ideal, assessment format. Conversely, from this same perspective, the more open-ended assessments that have been advanced as an alternative to multiple-choice formats will seem like a rather imprecise and inefficient measure of knowledge transfer.

Of course, association-level outcome measures are well-suited to many assessment tasks. Such items are generally answered quite quickly, allowing tests that cover a broad representation of the domain. By creating large pools of such items, random collections of such items provide assessments that are not biased toward any particular curriculum. When coupled with sophisticated psychometric techniques such as item response theory (IRT), empiricist assumptions about knowledge afford much of what the modern educational testing industry has to offer. IRT makes it possible to model the relative difficulty of specific items, and the relative proficiency of individuals relative to those items. This makes it possible to accurately predict the likelihood that an individual who demonstrates a certain level of proficiency on a broader pool of items will correctly answer a specific item. Once the relative difficulty of items has been established for a particular pool of students, items can be substituted as necessary to create new secure forms that efficiently and reliably compare students' familiarity with academic domains of knowledge. As long as one assumes that the items included on such tests are presumed to be an essentially random sample of the associations that make up the domain, such assessments represent a valid assessment of the transferable associations an individual possesses. In terms of content coverage, tests using such items are often assembled with a range of ability in mind. For each content area, a pool is assembled with items appropriate to that range in difficulty, and items are randomly selected from that pool in order to create a test that represents the content domain.

***Examples***

Conventional assessment practices and typical standardized achievement tests provide a wealth of examples of assessments of transferable knowledge that are consistent with empiricist assumptions. In many cases, these assumptions are implicit and unacknowledged (NRC, 2001b). Some of the best examples of assessment practices that explicitly acknowledge empiricist assumptions are those that follow from behavior-analytic perspectives (e.g., Fredrick, Dietz, Bryceland, & Hummel, 2000). For example, direct instruction methods directly teach very precise behavioral associations from a domain in very specific sequences (e.g., Carnine, Silbert, & Kameenui, 1997; Stein, Silbert, & Carnine, 1997). Following structured opportunities to practice those skills or concepts, students are directly assessed and given further instruction or opportunity to practice as needed. In the same vein, the artificial intelligence tutors associated with empiricist information-processing perspectives present students with very precisely sequenced cognitive associations from domains, and immediately assess mastery before allowing students to move on (e.g., Anderson, Boyle, & Rieser, 1985)

One of the goals of the aforementioned GenScope assessment effort was comparing the impact of the GenScope curriculum and conventional text-based curriculum on the relevant sections of a high-school graduation test. In transfer terms, a test was needed to measure the extent to which learning with GenScope transferred to an assessment that could also be validly administered to a non-GenScope comparison class. It was important to avoid bias (random or systematic) in favor of GenScope or a particular comparison environment. This is most readily accomplished by using a set of items that are drawn at random from the broad body of knowledge generally agreed to represent the domain of introductory genetics. This required having a large pool of items from which to draw—a challenge that is ideally suited to empiricist assumptions and a multiple-choice item format. Such a test was created from a set of 70 released SAT II biology content test items that covered genetics. A quasi-random sample of 15 items representing a known range of difficulty was created by ranking the items in order of reported difficulty, and selecting every seventh item. Because a similar process is used to select the genetics items on the graduation test, the research test provided a valid estimate of transfer to those items. The random selection of items minimized the chance of bias toward any particular curriculum; the pool of items of known difficulty made it possible to create another form of the test of similar difficulty (by selecting another 15 items in the same fashion).[1] This process resulted in a test that was then used as a valid tool for comparing learning in GenScope with learning in the conventional environments that GenScope was designed to supplant or replace. This resulted in what will be characterized below as a *distal*-level

assessment. As such, the same features that made it valid for comparing learning across a broad range of learning environments made it quite insensitive to learning within any particular learning environment.

## A Rationalist Perspective on Assessment of Transferable Knowledge

Rationalist perspectives became the major focus of psychological research on learning in the 1970s. While there was a shift away from the explicitly Piagetian "stage-theory" models in the 1980s, rationalist models continue to be very influential in educational and cognitive psychology. It is important to note at this point that the distinction between empiricist and rationalist perspective is not synonymous with "behaviorist" and "cognitivist" perspectives. In key respects, the "cognitive revolution" merely replaced behavioral associations with lower-level cognitive associations. The subsequent emergence of schema theories (e.g., Shank & Abelson, 1977) was one indication of the emergence of distinctly rationalist modes of cognitive science. (See Case, 1996, for a detailed discussion of the shifts within each of the three perspectives.)

### *Knowledge as General Concepts and Abilities*

Best understood as the antithesis of empiricism, this family of perspectives views knowledge in terms of structures of information and processes that recognize and make sense of (i.e., "rationalize") symbols in order to understand concepts and exhibit general abilities. The human mind is seen as a unique organ whose function is to make sense of the world. From this perspective, knowledge of a domain consists of both general reasoning schemata as well as more domain-specific concepts and specific skills. From this perspective, when an individual demonstrates knowledge of something, he or she is presumed to do so by marshalling the various higher-level knowledge structures needed to construct a solution for (i.e., make sense of) the demands of the particular task.

### *Learning and Transfer as Acquiring and Using Conceptual and Cognitive Structure*

When one assumes that knowledge consists of relatively general concepts, learning is the process of constructing the mental structures that represent those concepts. Piaget described the processes of assimilation (where patterns in the environment are assimilated into existing knowledge structures) and accommodation (where those knowledge structures are modified when they are no longer appropriate to make sense of some new pattern that is perceived). Thus the learning is presumed to occur as

students construct new mental structures of all sorts. This included their refinement of very high-level structures used to solve very general problems as well as the construction of very specific knowledge structures that are more relevant to specific domains.

When learning is viewed in terms of general knowledge structures, transfer of knowledge is then analyzed in terms of those same structures. In order to determine whether students have acquired knowledge that transfers, they must be presented with new problems that require them to apply that knowledge; the extent to which those problems differ from a specific or presumed learning environment, the farther that knowledge is presumed to be transferring. In contrast to the more random process that follows from empiricist perspectives, doing so generally calls for systematic consideration of the knowledge structures that define the domain (e.g., Mislevy et al., 2002).

***Assessment as Evaluation of Reasoning and Understanding***

When knowledge is viewed in terms of general conceptual schemata, assessment must examine students' ability to employ those schemata on larger, extended tasks. Thus, the appropriate environment for assessing students' ability to transfer knowledge from the learning environment presents learners with new problems that require them to apply the higher-level knowledge structures constructed in the learning environment. Some types of the familiar open-ended "constructed-response" items assess higher-level knowledge, and they can be obtained, administered, and scored with modest effort and high reliability. Among more stridently rationalist alternatives are the open-ended "performance assessments" that generally involve some sort of inquiry activity. Some involve hands-on activities, but many just present paper-and-pencil problem scenarios that students are asked to solve. The distinguishing characteristic of this broad array of approaches is the need to give explanations, or *rationale*, for phenomena.

As stated above, specific views of transfer guide, rather than dictate, the type of assessment one should use to assess knowledge transfer. In the extreme, a rationalist perspective argues that the specific lower-level associations that respondents appear to use when answering multiple-choice assessments are merely epiphenomenal artifacts of the item format. Instead, respondents are presumed to be using high-level schema to construct a solution to the (rather unnatural) problem. This perspective seems to underlie the stance of some observers who flatly reject standardized multiple-choice tests as valid assessments of the knowledge transfer (e.g., Kohn, 2000). In contrast, more modestly rationalist perspectives underlie many current efforts to broaden assessment practice (e.g., NRC, 2001b). This perspective argues instead that the structure of multiple-choice items biases assessment practices toward lower-level forms of understanding. Reflecting

modern cognitive science perspectives, such considerations generally distinguish between different forms of cognitive schema (e.g., *declarative, procedural,* and *strategic*; Ruiz-Primo & Shavelson, 1996) and emphasize that some forms of assessment are better suited to some knowledge in their consideration of different forms of assessment.

***Examples***

The open-ended problem-solving assessments that follow most logically from rationalist perspectives are challenging to develop and use, particularly in the context of large-scale, high-stakes testing programs (Solano-Flores & Shavelson, 1997). Fortunately, many such assessments have been developed and are available for use by educators and researchers. In science, for example, the PALS website (*Performance Assessment Links in Science*; Quellmalz, Schank, Hinojosa, & Padilla, 1999) offers numerous items with scoring rubrics, examples of student work, and item difficulty data.[2] In mathematics, new forms of the *Balanced Assessment in Mathematics* are developed and equated each year by the Mathematics Assessment Resource Services (MARS, distributed commercially by CTB/McGraw-Hill). There are many collections of performance assessment marketed for classroom use that can be used for program evaluation, and many sources offer useful guidelines for creating performance assessments (e.g., Stiggins, 2001; Wiggins, 1998).

Because performance assessments are relatively time consuming to administer and score, they cannot sample a range of content in the same fashion as multiple-choice tests. As such, items need to be selected and interpreted with care, particularly when using control-group designs. From a stridently rationalist perspective, these are simply technical problems, and are addressed by thoughtful analysis of the higher-level knowledge structures that represent the desired learning outcomes (elaborated in Hickey & Zuiker, 2003).

The GenScope assessment effort produced a sophisticated open-ended assessment based on a comprehensive model of expertise in the domain of introductory genetics (Hickey, Wolfe, & Kindfield, 2000). The various problems in this assessment followed quite directly from a detailed analysis of the reasoning skills that characterize expertise in the domain. The assessment presented increasingly complex problems, all involving the same simple organism. Sets of items reflected the research-based distinctions regarding domain reasoning (e.g., *cause-to-effect* versus the more challenging *effect-to-cause*, and *within-generation* versus the more challenging *between-generations*). Including such items allowed the research team to characterize the degree of knowledge transfer in terms of these same reasoning abilities. This sensitivity made it ideal for fine-tuning formative assessments in the GenScope environment (Hickey, Kindfield, Horwitz, &

Christie, 2003). This same sensitivity also introduced bias in favor of the GenScope curriculum and against any comparison curriculum, making it a *proximal*-level assessment, as will be explained subsequently.

Because rationalist perspectives continue to be quite influential in cognitive and educational psychology, they continue to be a central focus of efforts to assess knowledge transfer. The "evidence-centered design" approach by Mislevy and colleagues (e.g., 2002) is one of the most well-developed examples of assessment design that is consistent with the perspective outlined above. This approach starts with a sophisticated research-based model of reasoning in a domain, and then uses subtasks and sophisticated statistical models to systematically assess student understanding. Other notable examples include the IMMEX project (Interactive Multimedia Exercises; e.g., Stevens & Palacio-Cayetano, 2003) and FACETS (Minstrell, 2000).

## A Socioculturalist Perspective on Assessment of Transferable Knowledge

Sociocultural perspectives, particularly situative cognitive theory (e.g., Greeno et al, 1998), are still developing. While rooted in Vygotsky's seminal work following the Russian Revolution, these perspectives were largely unknown in Western education circles until the 1970s. They began to achieve prominence in the 1980s, and have continued to evolve within both education and cognitive science. An essential aspect of sociocultural perspectives that is often overlooked concerns the relationship between the individual and the broader environment in which the individual operates. Both of the preceding perspectives made a clear distinction between the individual and the environment. Sociocultural perspectives on knowing and learning lead to a view of transfer that assumes a dialectical relationship between the individual and the environment, in terms of the ongoing relation between the changing individual and the changing social context (Beach, 1999).

### *Knowledge as Distributed Cognition*

Sociocultural perspectives view knowledge as a cultural entity that is distributed across the physical and social environment in which that knowledge is developed and used. Thus an individual's knowledge of a domain is distributed across the people, books, computers, classrooms, worksheets, and so on that were present in the context in which the knowledge was learned and will continue to be used. Because knowledge is assumed to originate in the interaction of the social and material world, it is presumed that the social and material world is a fundamental part of that knowledge.

A fundamental assumption of this perspective is that participation in the use of knowledge changes the nature of that knowledge. This means that knowledge resides neither in the mind of knowledgeable individuals nor in the environment waiting to be derived "whole cloth." Rather, knowledge is "stretched across" the social and physical contexts of its use (Cole, 1991; Pea, 1985). Put another way, knowledge originates in the interaction of social and material worlds, and resides in socially defined tools and ways of interacting (Lave & Wenger, 1991).

From this perspective, knowledge is inextricably bound to the context of its use. A knowledgeable individual is one who participates successfully in sociocultural rituals and uses socially defined tools—what might best be called *knowledge rituals* and *knowledge tools*, or inclusively, *knowledge practices*. A student's knowledge of a domain such as introductory genetics includes the physical and social context in which that knowledge was developed. This knowledge in turn reflects the larger scientific community in which genetics is studied, as well as the community of learners in the particular classroom context or other environment in which that individual is learning about genetics. Reflecting an inherently contextualist worldview, every aspect of the context in which understanding of genetics is developed, is, at some level, part of that understanding. From this perspective, "knowledge" is represented in the regularities of successful activity. This regularity is possible because the "knowledgeable" individual has become attuned to the constraints (that bound participation) and affordances (that scaffold participation) of the environment in which successful activity occurs. This means that participants in knowledgeable activity are increasingly able to use physical and social tools to maximize successful participation and overcome the limitations of individual human minds.

The essential and controversial aspects of these perspectives concerns the distinction between *internalization* and *participation*. To many whose views are rooted in modern rationalist perspectives, sociocultural perspectives are understood in terms of learners' participation in the social construction of knowledge, which is subsequently internalized. Such views are consistent with what Lave (1991) labeled "cognition plus" and what Rogoff (1998) labeled "social influence theories." In these more modestly sociocultural theories, the social and cultural world is analyzed as a network of factors that shape and influence the development of internal knowledge structures and beliefs. Our characterization of a more uniquely sociocultural view of transfer reflects prior experience that clearly discriminating between rationalist and sociocultural views of transfer has practical value for improving assessment and testing practices (Hickey & Zuiker, 2003).

***Learning as Increasing Participation***

When knowledge is viewed as distributed across the social and physical context in which it is developed, learning is characterized as regular and successful participation in knowledge practices. Through this participation, individuals strengthen their respective ability to further participate in this activity. As the individual participates in a knowledge practice, the individual moves on an inbound trajectory from *legitimate peripheral participation* toward the center, toward *full participation* in the co-construction of that community's practice. According to Lave and Wenger, "legitimate peripheral participation is proposed as a descriptor of engagement in social practice that entails learning as an integral constituent" (1991, p. 35). The term "peripheral" is a positive term in that "the partial participation of newcomers is by no means 'disconnected' from the practice of interest" (Lave & Wenger, 1991, p. 37). The term *full participation* (rather than central or complete participation) was intended to do justice to the varying degrees of community membership. Learning, then, is not defined as acquiring isolated facts or as individuals making sense of the world. From this perspective, learning is a process of increasingly rich participation in communities of practice. For example, if the teacher is lecturing to the students on rote facts about the domain and then giving a quiz, learning is represented by the students' increasing ability to participate in that community of learners; similarly, if students are conducting laboratory experiments or using computer simulations, learning is represented by their increased ability to participate in that community as well.

***Transfer in Terms of Constraints and Affordances***

From this perspective, one must consider the constraints and affordances that support activity in the learning environment and in the transfer environment, and then consider "transformations" that relate to a given pair of learning and transfer situations. For transfer to occur, some constraints and affordances must be the same (be "invariant") across both situations, and the learners must learn (become "attuned" to) these *invariants* in the initial learning situation (Greeno, Smith, & Moore, 1993). From this perspective, many everyday transfer failures occur because curricular routines support learning of the "variant" aspects of a learning environment over the invariant aspect (cf. Greeno et al., 1998).

The unique implications of sociocultural perspectives for transfer are best revealed in conceptualizations that clearly break from prior conceptualizations. Beach argues that the entire metaphor of transfer should be left behind, arguing that the very notion confounds conceptual tools for understanding transfer with the phenomenon itself. Instead, Beach advances the notion of *consequential transitions*, involving "a developmental change in the relation between an individual and one or more social activities" (1999, p.

114). This represents a major shift of focus from prior individually oriented views of transfer. It allows, for example, the distinction between transitions between preexisting social activities (such as a unidirectional *lateral* transition from one course to another course) and transitions involving the creation of a new activity (such as the *mediational* transitions involving simulations where the collective participation of individuals supports a unique activity that bridges participation in activities that are developmentally prior and activities that are developmentally subsequent).

***Assessment of Participation***

In contrast to empiricist and rationalist perspectives that support a clear delineation between a learning environment and a transfer environment, sociocultural perspectives support the assessment of knowledge transfer within what would conventionally be defined as the learning environment. If the new knowledge that results from learning resides in the participatory rituals of the community (rather than in the minds of individual participants), assessment of knowledge transfer considers the way that rituals are adapted and appropriated by those participants. Because this participation necessarily changes that knowledge, learning is represented by the collective change in the participants and the physical and social context that supports their participation. This is why the *event*, rather than the individual, is the primary unit of transfer from a sociocultural perspective. As such, the use of interpretive event-based methods such as discourse analysis (e.g., Gee & Green, 1998) are ideally suited for assessing transfer from this perspective.

From this perspective, most interpretive analyses of learning conducted from a sociocultural perspective can be understood as the assessment of "knowledge transfer." Put differently, one can argue that the new knowledge that is seen when socioculturally oriented researchers document the emergence of domain-specific practices in communities of learners is evidence of knowledge "transfer." This applies, for example, to the emergence of "sociomathematical norms" in the analyses of mathematics classrooms by Cobb, Stephan, McClain, and Gravemeijer (2001). In our opinion, this realization is actually one of the major challenges in defining a sociocultural model of the assessment of knowledge transfer, and one of the key reasons that sociocultural views have had such a limited influence in mainstream considerations of assessment. It explains why, for example, conventionally minded observers might interpret the sociocultural vision of assessment outlined by Delandshere (2002) as more about studying "inquiry" than about the practices of "assessment." We contend that this confusion is the result of competing characterizations of the functions of assessment, which partly follow from competing views of knowing and learning. Our framework attempts to address this confusion by including

*level* and *function* as the second and third dimensions. As is shown below, sociocultural views of knowing and learning appear most well suited for the more proximal levels of assessment (i.e., *immediate* and *close*), and for largely *formative* functions.

***Examples***

The preceding paragraph highlights the challenges in identifying ideal examples of assessments of knowledge transfer that embody uniquely sociocultural assumptions about knowing and learning. Some theorists embrace an eclectic/hybrid approach that combines rationalist and sociocultural perspectives. For example, Schwartz, Bransford, and Sears (Chapter 1, this volume) advance such an approach to assessment that that they call *preparation for future learning* (PFL); this perspective is then presented as an alternative to the conventional empiricist view (which they characterize in terms of its *direct application,* or DA, view of transfer and the corresponding *sequestered problem-solving,* or SPS, methodology). This view of transfer has proven influential in the considerations of assessment by other theorists (e.g., Kolodner, Grey, & Fasse, 2003). In order to assess the transfer of students' participation in scientific inquiry in an innovative science curriculum, Duschl, Ellenbogen, and Erduran (1999) videotaped triads of participants while collaboratively critiquing a "standardized" science fair project, and then systematically coded those videos in terms of the quality of argumentation in that transfer context.

In the GenScope project, a great deal of effort was devoted to conceptualizing assessments of knowledge transfer from a sociocultural perspective. One hybrid approach that emerged was the quantitative coding of each turn of conversations during formative feedback sessions on a five-point scale ranging from *disruptive* to *scientific argumentation* (Schafer, Hickey, & Zuiker, 2003); another more distinctly sociocultural analysis (Zuiker & Hickey, 2004) used discourse-analytic methods to interpret the emergence of "epistemic stance" as students negotiated norms for scientific argumentation (e.g., Ochs, 1996). In the end, most of the socioculturally inspired "assessment" took the form of print- and video-based guidelines that were designed to help students and teachers informally assess the domain discourse that occurred as they enacted the GenScope investigations and engaged in "feedback conversations" around classroom assessments.

This points to a potentially more appealing option for sociocultural assessment of knowledge transfer. This is the use of multiple forms of assessment at different levels. As described in more detail below, it may make most sense to use interpretive methods to examine the sociocultural transfer of knowledge rituals within a learning environment, while using more conventional measures to assess how the knowledge constructed in those events transfers to individual performance in a conventional transfer

context. As was elaborated above, views of knowing and learning guide one's interpretation of assessment practices, rather than dictating assessment methods. This insight is crucial to appreciating what may ultimately be the most useful aspects of sociocultural views for assessing knowledge transfer. Greeno and colleagues (1998) argued that sociocultural views offer a "higher-order synthesis" that can be used to reconcile the strengths and weaknesses of prior empiricist and rationalist perspectives. In this view, one can treat *all* individual outcome measures as "peculiar" knowledge rituals that are necessary for certain purposes. From this perspective, an individual's score on a performance assessment or a multiple-choice test is seen as evidence of their success in participating in a particularly constrained form of domain discourse (Hickey & Zuiker, 2003; NRC, 2001b). We will return to several of the issues raised above, highlighting our fundamental argument that any systematic effort to assess knowledge transfer needs to take into account all three of the dimensions that make up our framework.

## MULTIPLE LEVELS OF ASSESSMENT AND TRANSFER

The second dimension in our framework concerns *levels* of assessment. Ruiz-Primo and colleagues (2002) point out that different types of assessments represent increased "distance" from the enactment of a particular curricular activity. This continuum is akin to the distinction between "near" and "far" transfer that is central to most of the chapters in this volume, in that it represents distance from the context in which some learning has occurred, and the context in which transfer of that learning is being assessed. As shown in Table 7.2, and following directly from Ruiz-Primo and colleagues, we identify five discrete points on the continuum of distance as it pertains to assessment: *immediate* (e.g., informal observation or artifacts from the enactment of a particular lesson), *close* (e.g., embedded assessments and semi-formal quizzes following several activities), *proximal* (e.g., formal classroom exams following a particular curriculum), *distal* (e.g., criterion-referenced achievement tests such as required by NCLB), and *remote* (broader outcomes measured over time, including norm-referenced achievement tests such as the ITBS).

These distinctions are important because they frame any discussion about the validity of assessments at each level as evidence of transferable knowledge. While the actual number and nature of levels might change for different purposes, what constitutes valid assessment of knowledge transfer varies across them. For example, criterion-referenced and norm-referenced achievement measures are too far removed from a specific learning activity to accurately assess knowledge transfer. Thus, for example, if one wants to assess the transfer of learning from two different versions of a spe-

cific curricular activity, the knowledge that is assessed in conventional achievement tests is too broad and too removed from the curricular routine to detect any meaningful differences. Conversely, these same features are what make achievement tests valid for some kinds of large-scale comparisons of entire curricula (because they may preclude random or systematic bias toward a particular one.

An important aspect of levels is the *orientation* of assessments at each level. In order to better characterize discrete points on the continuum of distance, we have identified what appears to be the most appropriate focus of assessments at each level. As shown in the second column of Table 7.2, this ranges from specific events at the *immediate* level to national achievement at the *remote* level. In this regard, we have found that the assessment of knowledge transfer at each level and corresponding orientation can be best understood in terms of Lemke's (2000) notions of *timescale*, as shown in the third column of Table 7.2. We contend that timescale is relevant to the assessment of knowledge transfer because the different competencies that different assessments aim to measure are "timescale-specific." At the immediate level, knowledge transfer can be assessed within the discourse that defines specific learning events over a period of a few minutes. This is very different from criterion-referenced achievement tests at the remote level, which are typically oriented toward state-level content standards and measure knowledge transfer across school semesters, a timescale of roughly months. By the same token, norm-referenced tests at the remote level must function on a timescale of years in order to provide percentile rankings that are interpretable from one year to the next. To the extent that this is the case, learning and transfer are also timescale-specific. As will be elaborated in the next section, an understanding of timescales is also crucial to understanding and balancing the formative and summative potential of each of the levels.

## MULTIPLE FUNCTIONS OF ASSSESSMENT

The final dimension in our framework concerns *functions* of assessment. Most readers will be familiar with the distinction between *summative* assessment carried out to provide evidence of prior learning and *formative assessment* carried out to advance learning. Assessment theorists argue that this is a crude distinction that implies a false dichotomy between formative and summative functions (NRC, 2001b; Shepard, 2000). In fact, most "summative" assessment practices ultimately aim to serve some formative goal, and all "formative" assessments must include some summative functions to succeed. For example, while the 4-year administration cycle of the National Assessment of Educational Progress (NAEP) makes it the epitome of "sum-

**Table 7.2. Five Levels of Assessment**

| *Level (Example)* | *Primary Orientation* | *Time Scale Relationship to Curricula* | *Optimal Formative Function* | *Optimal Summative Function* | *Prototypical Domain Knowledge Representation* | *Prototypical Assessment Format* |
|---|---|---|---|---|---|---|
| IMMEDIATE (artifacts from the enactment of the curriculum) | Specific Events | Minutes | Guide and refine enactment of specific curricular routines. Informally advance participation domain discourse. | Informally assess whether routines are enacted as intended within specific activities. | Enacted knowledge practices, discourse during a specific curricular routine. | Event-oriented *observations*; guidelines and exemplars for classroom discourse. |
| CLOSE (semi-formal classroom assessment) | Specific Activities | Days | Give all students more robust understanding of the topic. Support semi-formal remediation for specific students and/or topics. Guide the refinement of specific activities. | Assess whether students can engage in discourse around the content in specific activities. | Semi-formal representation, discourse around representations similar to the curricular routines. | Activity-oriented *quizzes* and discursive formative feedback. |
| PROXIMAL (formal classroom assessment) | Entire Curricula | Weeks | Help all students transfer new knowledge to more formal representation. Support formal remediation for specific students and/or topics. Guide refinement of curricular sequence. | Assess whether students learned intended content in specific curriculum. | Formal representation of the concepts covered in the curriculum. | Curriculum-oriented *exams* and conceptual formative feedback. |
| DISTAL (criterion-referenced external tests) | Regional or National Content Standards | Months | Help teachers and researchers see if students can use knowledge in formal high-stakes context. Guide refinement of entire curriculum accordingly. | Assess whether students meet specific criteria, compare curricula in this regard. | Formal representation of associations drawn from the standards. | Criterion-referenced tests carefully aligned to content standards. |
| REMOTE (norm-referenced external tests) | National Achievement | Years | Help researchers and policymakers see if curricular standards and reforms are effective. | Impact of broader changes to curriculum and standards. | Formal representation of associations drawn from national samples of achievement. | Norm-referenced tests. |

mative" assessment, NAEP's formative value for long-term educational policy becomes apparent over the timescale of generations of learners. Conversely, even informal classroom observations necessarily include some summative comparisons in order to meet more formative goals, such as guiding instruction and remediation. The summative functions are less obvious in part because such comparisons occur within a timescale that is almost instantaneous. A related observation is that particular forms of assessment often serve multiple formative and summative functions. For example, teachers can also use informal classroom assessments to directly advance student learning as well as refine specific curricular routines; administrators often use criterion-referenced tests to promote students, compare schools, and guide large-scale curricular reforms. Our first point is (1) there are a wide range of formative and summative functions of assessment, (2) these functions differ fundamentally across levels, and (3) there are multiple formative and summative functions within levels.

Our second point, to be elaborated below, is that the summative functions of assessment often undermine formative functions. This typically occurs when summative concerns over "evidential validity" and experimental control override more formative concerns over "consequential validity" and generalizability (Shepard, 1993; Wiliam & Black, 1996). This is relevant to the broader study of knowledge transfer because most considerations of transfer (including the chapters in this volume) also have formative goals for improving education and learning. Therefore, it is possible that the summative concerns in the measurement of transfer may also undermine the broader formative intent of transfer research. To the extent that this is true, our efforts to balance formative and summative functions of assessment should apply to considerations of knowledge transfer in general.

## Balancing Formative and Summative Assessment Functions

Assessment is ultimately about *feedback*. The value of feedback from assessment is well documented. A review of research on classroom assessment by the National Research Council (2003) reviewed 40 experimental and quasi-experimental studies previously catalogued by Black and Wiliam (1998) with effect sizes on achievement ranging from .40 to .70. These findings, along with 250 other books and articles, led Black and Wiliam to conclude "we know of no other way of raising standards for which such a strong *prima facie* case can be made on the basis of evidence of such large learning gains" (p. 19).

Black and Wiliam's (1998) landmark review summarized concerns that assessment is effective only to the extent that information it generates is actually used to advance learning and instruction. Whether it is used for promoting or selecting individuals, evaluating schools, refining curriculum, or directly advancing knowledge, the value of assessment is defined by the usefulness of the feedback it provides. This insight is central to most prior considerations of the formative value of assessment (e.g., Black, 1993; Crooks, 1988; Natriello, 1987; Pellegrino et al., 1999; Sadler, 1989; Stiggins, 2001; Torrance & Pryor, 1998; Turnstall & Gipps, 1996). While this insight is most typically associated with classroom assessment practices, it is also central to large-scale assessment reform efforts (e.g., Herman, 1997). And while some argue that recent assessment-driven reforms such as No Child Left Behind are more about politics than about pedagogy (e.g., Kohn, 2004), we contend that consideration of these issues will minimize the negative consequences of accountability practices and maximize their formative potential.

The NRC (2001a) classroom assessment report reviewed the evidence that much of the potentially useful information from classroom assessment is unused or is used in ways that undermine learning. At a minimum, nonformative assessments consume valuable instructional time; excessively summative assessments can lead to corrosive comparisons of past performance and cause some low achievers to stop trying, or quit school altogether (Amrein & Berliner, 2003; Kelleghan, Madaus, & Rackzek, 1996). Another issue is that different forms of assessments (e.g., multiple-choice vs. open-ended, informal vs. formal, embedded versus nonembedded, criterion-referenced vs. norm-referenced, etc.) provide feedback that is more useful for some formative functions than others. This explains limitations of the ubiquitous "test" prep practices being aggressively and successfully marketed to schools as "evidence-based" solutions to student achievement (e.g., Yesseldyke & Tardrew, 2002). Closer scrutiny shows that, at best, such practices yield modest gains on specific tests, with little or no evidence of generalized student learning beyond the targeted test. This is actually not surprising given that the feedback in many test-prep programs just tells students which was the correct choice. Even programs that offer more detailed feedback on practice forms of standardized tests are hard-pressed to support meaningful student learning. This is because an item stem only loosely relates to most of the answer choices and any one item only loosely relates to other items. Subsequent learning is then structured around the relatively disconnected associations represented by the various items. One increasingly common informal test-prep practice, directly instructing students on recently released items, can actually decrease test scores (when test-makers avoid similar items on the subsequent version). Conventional achievement tests are also problematic for curricular refinement and

remediation because they cannot be closely related to classroom instruction. They do not relate to everyday classroom activities and they cannot provide feedback useful to teachers refining their curriculum or provide remediation for individual students or on individual topics (Commission on Instructionally Supportive Assessment, 2001).

## Balancing Functions between *and* Within Levels

Most prior considerations of the balance between formative and summative assessment functions have focused on tensions between assessment levels:

> The lack of coherence among the different levels of assessment within the system often leave the teachers, schools, and districts torn between mandated external testing policies and practices, and the responsibilities of teachers to use assessment in the service of learning. (NRC, 2001a, pp. 72–73)

> …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)

It seems that this focus on alignment between levels is partly responsible for the widely held dichotomy between formative assessment and summative assessment. We contend that this focus has obscured the way that summative functions undermine formative goals *within* levels of assessment. For example, many treatments of classroom assessment focus on the feedback provided to teachers (e.g., Anderson, 2003). As such, their impact on student learning is relatively *indirect*, via promotion, remediation, and refinement. From the students' perspective, there is nothing inherently formative about classroom assessment when it is used for these purposes; as such, these more summative functions may interfere or undermine formative goals (e.g., by creating corrosive levels of competition, focusing attention on "fairness" rather than understanding, etc.).

Balancing formative and summative functions within levels is particularly important when focusing on assessments for *directly* advancing student learning, via feedback provided directly to learners (e.g., Duschl & Gitomer, 1997; Hickey et al., 2000). As detailed below, informal classroom assessments are ideal for the largely formative function of directly advancing student knowledge. However, some summative functions are needed to give learners accurate feedback about their own knowledge in order to guide their participation, and are sometimes needed to motivate initial

engagement in the assessment as well. In such cases, care must be taken to define the right degree of summative function.

Balancing functions within levels is particularly crucial when a single assessment is used to serve multiple formative and summative functions. For example, semi-formal classroom assessments (such as the close-level *quizzes* detailed below) are likely to be used for both directly advancing student knowledge and guiding the refinement of specific curricular routines. However, the ideal level of summative function for the first formative goal might be insufficient for the second. Given the likelihood of having multiple summative and formative goals for a single level, this seems like a particularly important issue. As shown in Table 7.2 and elaborated in the next section, we address this issue by attempting to define optimal formative *and* summative functions for each level of assessment.

### Optimal Formative and Summative Functions by Level

Summaries of our current understanding of optimal formative and summative functions by level are listed in the fourth and fifth columns of Table 7.2. These definitions follow directly from the orientations and corresponding timescale that are listed in the second and third columns, and reflect the idea outlined in the preceding paragraphs. While it is important to reiterate that the five levels are points on a continuum, we also contend that defining discrete levels is crucial for balancing formative and summative functions. Working with at least three such adjacent levels at a time (e.g., close, proximal, and distal) makes it possible to use formative feedback at the first level to improve outcomes at the second level (maximizing "consequential" validity), while preserving the summative value of the third level (maximizing "evidential" validity). As described in the next section, we contend that this balancing is further facilitated when different forms of knowledge representation are explicitly acknowledged and appropriately used as an interpretive base at each level.

## COMBINING THEORY, LEVEL, AND FUNCTION TO FACILITATE ASSESSMENT OF KNOWLEDGE TRANSFER

The examples of assessments at each level described above alluded to a systematic relationship between theories of knowing and learning and levels of assessment. Specifically, the claim we advance here is that the immediate-level observations of curricular enactments are more consistent with sociocultural assumptions about knowing and learning, the proximal-level problem-solving classroom assessments are more consistent with rationalist

assumptions, and the remote-level norm-referenced achievement tests are more consistent with empiricist assumptions. Our framework aims to advance assessment practice by arguing that there are actually fundamental relationships between theories and levels, and that the diverse educational goals of assessment will be better met if assessment practice recognizes and embraces them. This is especially important in avoiding persistent misunderstandings by educators, policymakers, the public, and the media arising from the misuse of assessment results at one level to make evaluative judgments of educational outcomes associated with another level.

Our initial characterization of the relationship between theories and levels is detailed in the sixth column of Table 7.2. For pragmatic purposes, we choose to characterize this relationship in terms of the prototypical or "best-fitting" manner for representing knowledge at each level. These are initial characterizations and will continue to need refinement. For now, we propose that assessment practices at the immediate, proximal, and remote levels prototypically need to embrace sociocultural, rationalist, and empiricist assumptions, respectively, while the close and distal levels are most likely to employ hybrid sociocultural/rationalist and rationalist/empiricist assumptions.

Implicit in these characterizations is the idea that transfer across different forms of knowledge representations is unidirectional. While we have yet to prove this assumption experimentally, the prototypical assessments in the next section emerged from efforts to maximize the consequences of formative feedback at one level for summative performance at the next level (e.g., Hickey et al., 2003). These efforts have convinced us that the "cultural" knowledge that results from meaningful participation in domain discourse transfers readily "cognitive" representations of that knowledge, and transfers moderately to "behavioral" representations of that knowledge—*but the inverse is usually not true.* To use the examples below, this means that formative feedback on proximal-level performance assessments will likely help students recognize correct associations on distal-level norm-referenced tests, but that formative feedback on norm-referenced tests will likely *not* help students solve problems on performance assessments (and might actually hamper participation in classroom discourse).

## Prototypical Assessments by Level

We close our consideration by outlining an initial set of "prototypical" assessments at each level that have emerged in prior and continuing assessment research efforts. These prototypes are summarized in the seventh column in Table 7.2 and detailed below. Included in these descriptions are an

initial set of references to existing research literature that seem useful for guiding efforts at that level.

***Immediate-Level Event-Oriented Observations***

Directly following from the cultural characterization of knowledge outlined in the sociocultural perspective, knowledge at this level in our approach is represented by the enactment of knowledge practices in specific curricular routines. Specifically, these refer to students' engagement in the discourse practices of the domain during the actual lesson. By discourse we mean conversations, as well as any interaction with the symbols and signs of the domain. Ideal assessment at this level consists of informal observation during curricular routines. Because teachers and students can directly observe discourse, formative feedback can also directly and immediately enhance it. This means that the immediate level is ideally suited for directly advancing student knowledge. This feedback is also useful for providing teachers with useful guidance regarding the structure of the individual curricular routines, and guiding their refinement. As such, assessment at the immediate level should be almost entirely formative. Including formal summative components (such as assignment of grades) is likely to undermine formative value.

Because they are oriented to specific events and represent a timescale of minutes, immediate-level assessments yield feedback that is useful for fine-tuning curricular routines as they are being enacted (see, e.g., Wenk, Dufresne, Gerace, Leonard, & Mestre, 1997). Good teachers already do this. We contend that helping all teachers *and students* do so more successfully is an ideal aim of immediate-level assessment practices. This is best accomplished via guidelines for teachers and students that structure classroom discourse. This actually characterizes much of the work in formative assessment that has been inspired by sociocultural perspectives. Relevant considerations in the research literature include Delandshere (2002) and Torrance and Pryor (1998). Both Turnstall and Gipps (1996) and Duschl and Gitomer (1997) provide helpful taxonomies of characteristics of worthwhile domain discourse in the context of formative assessment.

In the GenScope project, immediate-level efforts focused on including additional genetics content within the curricular materials that guided students' computer-based explanations, and coaching the teacher to coach the students to discuss those topics when they were conducting the investigations. New considerations of immediate-level assessment are taking place in an ongoing pilot project in elementary mathematics (Hickey, 2004). This project is expanding the guidelines for teacher observation in existing statewide lesson plans using the *Understanding by Design* framework advanced by Wiggins and McTeague (1999). The project is also developing simple Web-based dramatized video clips of lower-quality and higher-qual-

ity enactments of those lessons, having students use them to informally self-assess their own enactments.

***Close-Level Activity-Oriented Quizzes and Discursive Feedback***

Prototypical close-level assessments should embody both sociocultural and rationalist assumptions about knowing, learning, and transfer. As we now envision it, ideal close-level assessments consist of semi-formal representations of domain knowledge that are similar to the ones students encountered in the curricular context. This is a critical point, because students and teachers are presumed to have negotiated a shared understanding of these representations. This in turn provides the common knowledge base that is crucial for semi-formally assessing and further advancing this understanding. Close-level assessments should also maintain a semi-formal administration context, presenting only the minimal-level external accountability (e.g., grades, "participation points," etc.) needed to motivate initial engagement. This is critical because excessively summative contexts will direct student attention to prior performance and corrosive comparisons, and away from advancing their current understanding. This appears necessary to allow a "local" accountability to flourish, where students hold themselves and each other accountable for their understanding and participation in authentic domain discourse when engaging in formative feedback. In summary, ideal close assessments should let students individually solve domain problems, and then engage in authentic domain discourse in the context of formative feedback. The formative feedback they provide is ideal for directly advancing student knowledge, and is also quite useful for guiding semi-formal remediation by teachers and for refining specific curricular routines.

While they may not be characterized as we have, there are numerous examples in the research literature of assessments that have been (or might be) used this way. Most obviously is the *Preparation for Future Learning* assessments described by Schwartz, Bransford, and Sears (Chapter 1, this volume). A particularly well-known example is White and Frederiksen's (1998) *reflective assessment* model that was refined in the context of a computer-based inquiry-learning environment in physical science. After students generated new situations and experimental plans that they could use to test and improve their conceptual model of physics, White and Frederiksen would have them collaboratively judge their efforts against a rubric that detailed the criteria for authentic scientific inquiry. Other assessment practices that are consistent with our ideal are found in the groundbreaking work described in Wolfe, Bixby, Glenn, and Gardner (1991) and in some of the formative assessment practices reviewed in Black and Wiliam (1998).

A great deal of the effort in the GenScope project was concerned with refining close level assessment and formative feedback, though without labeling it as such (e.g., Hickey et al., 2000). The project refined sets of semi-formal quizzes that used the same traits and organisms as the GenScope software. After individually completing the quizzes, students' assessments included using a "learner-oriented" formative feedback rubric. These rubrics offered detailed explanations of the reasoning behind each item, using authentic academic language (vs. "everyday" vernacular). Unlike conventional scoring rubrics, they also include details that are not "technically" necessary. Most importantly, the rubrics did not directly state the correct answer; rather, students had to collaboratively make sense of the rubric in order to determine the most suitable answer (Hickey et al., 2000). Students used their completed assessments and the rubrics to discuss their understanding of the assessed topics during carefully orchestrated "assessment conversations." In the 20-hour GenScope curriculum, underachieving ninth graders made dramatic gains on distal-level and remote-level outcomes after completing three such assessments and three hour-long formative feedback conversations (Hickey, Zuiker, & Kindfield, 2004). These methods are being further refined in the mathematics pilot project mentioned above and in studies of NASA-funded science curriculum currently underway (Hickey, 2003).

***Proximal-Level Curriculum-Oriented Exams***

Assessments at the proximal level are oriented toward specific curriculum and represented on a timescale best measured in weeks. We contend that ideal proximal-level assessments follow most directly from a rationalist perspective, particularly as embodied in the modern cognitive perspective that is reflected in many current calls to broaden assessment practice. Ideal assessments in this regard should provide a formal representation of the concepts covered in the specific curriculum. As embodied in the conventional end-of-semester exam, the primary function of feedback from proximal-level assessments should be ensuring that students understand the content of an entire curriculum. This information is essential for local accountability (i.e., grading), revising entire curricular sequences, and supporting formal remediation for specific students and topics. This information also appears useful for supporting additional feedback to directly advance student learning, especially in helping students learn how their less formal understanding of the curricular domain translates to a more formal assessment context.

It is not yet clear what is the ideal assessment format for this level; it may be that a range of formats (e.g., open-ended, short-answer, or multiple-choice) are variously appropriate depending on the nature of the knowledge. The important thing is that proximal-level assessments cover all of

the content that students should have mastered in the particular curriculum. While open-ended assessments might be a more accurate judge of mastery, they may be too time-consuming in many situations. Some of the assessments reviewed in Black and Wiliam (1998) appear appropriate for proximal-level goals. From our perspective, many of the assessment practices from prior large-scale assessment reforms (such as described in Herman, 1997) might actually have been more appropriate for use in this context (rather than in the more distal-level setting for which they were designed).

In the GenScope project, the proximal-level assessment consisted of the open-ended performance assessment described above. In current efforts, formal classroom assessment consists of multiple-choice items that have been "cherry-picked" from released state tests. Items are selected because they directly match the content of the curriculum, but assess that understanding in a formal test context. Items are selected to help ensure that core curricular goals are met. For example, we select items that reveal to students and teachers the logical "traps" or persistent misconceptions that item writers exploit to make difficult items for simple concepts. The exams are administered after an entire curriculum (covering one or more months of classes). Because the exams are closely aligned with the close-level assessments, their feedback is ideally suited to refining that aspect of the curriculum. We have also begun experimenting with providing students with formative feedback rubrics similar to the ones provided for the close-level assessments. Initially, it appears that completing the close-level quizzes and feedback conversations prepares students to take full advantage of the formative feedback on the exam, and keep the more salient summative functions of the exams from undermining their easily overlooked formative potential (Taasoobshirazi, Zuiker, & Hickey, 2004).

***Distal-Level Criterion-Referenced Tests***

External achievement tests used to determine whether students have mastered the content in specific standards are the epitome of distal-level assessments. Such tests are the centerpiece of No Child Left Behind. While teachers are expected to have prior knowledge of the standards to which the items are aligned, it is crucial that teachers and students not have prior knowledge of the particular items, and that new items be regularly introduced to prevent compromise. This requires having pools of items for particular standards and of a known level of difficulty. These functions require highly structured formats given typical constraints of testing time, broad content coverage, and a census-testing model (all students receive the same test at a given point in time). Under such circumstances, multiple-choice items are ideal, and short-answer items are also useful. In some cases, such as writing tests, open-ended formats are necessary and generally

manageable. If the constraints are relaxed such as a matrix sampling model for item sampling and test construction that is typically combined with elimination of the need for individual student scores, a variety of testing formats may be used, including extended constructed response and performance assessment tasks. Such models have been used successfully in state achievement testing programs such as Maryland's MSPAP.

The highly formalized representation of domain knowledge and the summative administration context for most distal-level criterion-referenced tests yields feedback that is useful for (1) formally determining whether individual students or collections of students have acquired some particular familiarity with specific domains; (2) assessing the general success of specific teachers and schools in this regard, appropriately taking into account other factors contributing to variance in both score levels and score gains; and (3) evaluating the impact of various curricula and instructional programs. Distal-level assessments need to employ both rationalist and empiricist assumptions; the former help ensure that the feedback will be useful for evaluating the impact of curricula and instructional programs and perhaps suggesting areas in need of revision, while the latter support critical summative functions such as test score reliability and comparability.

In many settings, student scores from school-administered criterion-referenced content tests can be very useful. In our projects, we create proxies for such tests and use them to refine and evaluate our curriculum. For such proxy tests to provide valid and accurate estimates of actual impact, it is essential that such assessments be developed and administered in the same fashion as the operational tests. As described above, we used released SAT II items in the GenScope project. So long as teachers did not see the test (and curriculum developers did not refer to it), the test was a legitimate proxy for the corresponding parts of the graduation test (roughly one-ninth of the test).

***Remote-Level Norm-Referenced Tests***

Remote-level tests are quite similar to criterion-referenced tests, but they are standardized very carefully 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) is one of several such tests used to evaluate student achievement. A particularly useful feature of the ITBS and other such tests is the calibration 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. 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.

The stable, highly formal representation of domain knowledge, which may or may not be linked to national or state content standards, and the highly formal administration context provides feedback that is useful for comparing the achievement of many students over extended periods of time. As such, this feedback is also useful for evaluating the impact of efforts to finetune schools to maximize performance on criterion-referenced tests. It is a well-established fact that scores routinely rise over the first few years of performance on a state's criterion-referenced achievement test even though performance on other measures such as NAEP may remain stable (Linn, 2001). Interestingly, some of the controversy over the No Child Left Behind Act concerns evidence that norm-referenced scores are falling as a direct consequence of efforts to raise scores on criterion-referenced tests (Hoff, 2004; Markley, 2004; Schemo & Fessenden, 2003). Whether this performance tradeoff is important or not depends on the educational objectives one has in mind and the relative weight one places on different sources of evidence that are derived from multiple levels of assessment data. This also underscores the point that educational systems need coordinated systems of assessments that are articulated across levels in ways that reinforce three components: comprehensiveness, coherence, and continuity (see NRC, 2001b).

## SUMMARY AND CONCLUSIONS

In summary, this chapter has argued that the assessment of knowledge transfer in education will better meet its diverse goals by seriously considering the three dimensions that make up our framework and their interaction. We have argued that assessment practices must balance formative and summative functions, and that doing so requires recognition of the unique formative and summative functions of different assessment levels. We suggested that working with at least three such adjacent levels (e.g., close, proximal, and distal) is particularly useful because it makes it possible to finetune the formative functions of the first level using feedback from the second level, while preserving the evidential validity of the third level. We have also suggested that the different formative and summative functions of different levels are most likely to be accomplished by shifting from sociocultural to rationalist to behaviorist representations of knowledge across those levels. In doing so, we have argued that knowledge transfer across different ways of representing domain knowledge may be unidirectional. This provides a transfer-based psychological explanation of the limitations of "test prep" practices that provide formative feedback on standardized

achievement tests; it may also explain why many reform efforts involving performance assessments had little impact on learning or achievement.

The explication of our framework in this chapter was motivated in part by the enduring tensions over educational accountability (e.g., Linn, 2000, 2001). The recent No Child Left Behind Act has directed new attention to student assessment and generated substantial arguments among policymakers, researchers, and educators. This chapter does not attempt to resolve those arguments. Rather, it implies that their satisfactory and scientific resolution requires more careful consideration of the nature of school learning and the role that assessment plays as a measure of the transfer of that learning. We would argue that if policymakers and researchers are to use student assessments wisely to make scientific and policy arguments about "what works," they need to consider the assumptions that underlie varied and often competing forms of assessment. Our central conclusion concerning NCLB is that meeting its ostensible goal of substantial improvement in education and achievement for all students requires at least two levels of classroom-level assessment and useful formative feedback (e.g., close-level quizzes and proximal-level exams) that are systematically aligned to the relevant criterion-referenced tests that are required by NCLB. We further conclude that the true impact of these refinements should then be documented using distal-level norm-referenced achievement tests to ensure that the distal-level gains are not simply due to a narrowing of the curriculum. It remains to be seen whether the overall impact of these refinements can or should be systematically assessed using even more remote tests such as NAEP.

Efforts to address the issues raised above reflect appreciation of an even broader set of issues concerning the very nature of educational research. The NCLB legislation uses the term *scientifically based* over 100 times, and American legislators have initiated unprecedented efforts to specify what qualifies as scientifically based educational research (e.g., House Resolution 4875). Another report by the National Research Council (2002) attempted to clarify the issue, but other events suggest that the issue has become even more complex and contentious (Olson & Viadero, 2002). For example, the U. S. Department of Education's Strategic Plan (2002) argued

> The field of education operates largely on the basis of ideology and professional consensus. As such it is subject to fads and incapable of the cumulative progress that follows from the application of the scientific method and from systematic collection of objective information in policy making. We will change education to make it an evidence-based field. (p. 48)

This and numerous other policy documents have attempted to establish randomized experimental designs as the hallmark of "scientifically based"

educational research (United States Department of Education, 2002; Stanovich & Stanovich, 2003). In addition, new criteria for educational evaluation treat externally developed assessments (such as standardized achievement tests) as more objective, and therefore more scientific, compared to assessments developed by educators and researchers (United State Department of Education, 2003). As we have shown, theoretical assumptions spelled out in the first dimension strongly impact one's research questions about learning and transfer; these questions, in turn, impact selection of assessments and their use in research designs. Similarly, specific assessments are more appropriate for assessing particular types of learning, at particular levels, and for particular functions. In closing, we would argue that careful consideration of the relationships between transfer and student assessment, such as we have outlined in this chapter, are essential if we are to move beyond polemics about the quality of educational research and adequately debate the evidence appropriate for various research designs. Such debate demands careful analysis of the inferences about student achievement that various types of evidence can and cannot support.

## ACKNOWLEDGMENTS

This chapter is based on work that was supported by the U.S. National Science Foundation (Grant No. REC-0196225), the NASA-funded Center for Educational Technology at Wheeling (WV) Jesuit University, and the University of Georgia. The opinions represent those of the authors. Steve Zuiker, Nancy Schafer, and Marina Michael were instrumental in the refinement of the ideas in this chapter; Rachel Lewis and Dionne Cross provided substantial input on the preparation of this manuscript.

## NOTES

1. Illustrating the tradeoffs that occur when measuring transfer in this way, the test was actually biased against the GenScope curriculum. For example, it ended up including an item on Lamarkian reasoning (the common misconception that acquired characteristics, as in trimming a dog's tail, are inherited). The fact that acquired characteristics are never inherited is directly taught in most conventional curriculum (the common misconception that acquired characteristics are passed on to offspring), but was not directly addressed in the GenScope curriculum.
2. The PALS website is located at http:// www.ctl.sri.com/pals/

## REFERENCES


Amrein, A. L., & Berliner, D. C. (2003). The effects of high-stakes testing on student motivation and learning. *Educational Leadership, 60* (5), 32–38.

Anderson, J. R., Boyle, C. F., & Reiser, B. J. (1985). Intelligent tutoring systems. *Science, 228,* 456–462.

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

Beach, K. D. (1999). Consequential transitions: A sociocultural expedition beyond transfer in education. *Review of Research in Education*, 24, 124–149.

Bereiter, C., & Scardamalia, M. (1996). Rethinking learning. In D. R. Olson & N. Torrance (Eds.), *The handbook of education and human development*, (pp. 485–513). Cambridge, MA: Blackwell.

Black, P. J. (1993). Formative and summative assessment by teachers (1993). *Studies in Science Education, 21,* 49–97.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Bloom, B. S. (1976). *Human characteristics and school learning.* New York: McGraw-Hill.

Carnine, D. W., Silbert, J., & Kemeenui, E. J. (1997) *Direct instruction* reading (3rd ed.). New York: Prentice Hall.

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). Blackwell: Cambridge.

Cobb, P., Stephan, M., McClain, K., & Gravemeijer, K. (2001). Participating in classroom mathematical practices. *Journal of the Learning Sciences, 10*, 113–164.

Cole, M. (1991). On socially shared cognitions. In L. Resnick, J. Levine, & S. Behrend (Eds.), *Socially shared cognitions* (pp. 398–417). Hillsdale, NJ: Erlbaum.

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.

Commission on Instructionally Supportive Assessment. (2001). *Building tests to support instruction and accountability.* http://www.nea.org/issues/high-stakes/buildingtests.html

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58,* 438–481.

Delandshere, G. (2002). Assessment as inquiry. *Teachers College Record, 104,* 1461–1484.

Duschl, R. A., Ellenbogen, K., & Erduran, S. (1999, March). *Promoting argumentation in middle school science classrooms: A project SEPIA evaluation.* Paper presented at the annual meeting of the National Association for Research in Science Teaching, Boston.

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.

Fredrick, L. D., Dietz, S. M., Bryceland, J. A., & Hummel, J. H. (2000). *Behavior analysis, education, and effective schooling.* Reno, NV: Context Press.

Gagné, R. (1985). *The conditions of learning and a theory of instruction* (4th ed.). New York: Holt, Rinehart & Winston.

Gee, J. P., & Green, J. (1998). Discourse analysis, learning, and social practice: A methodological study. *Review of Research in Education* 23, 119–169.

Gipps, C. V. (1994). *Beyond testing.* London: Falmer.

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., Smith, D. R., & Moore, J. L. (1993). Transfer of situated learning. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction.* (pp. 99–167). Norwood, NJ: Ablex.

Greeno, J. G., and the Middle School Mathematics through Application Project. (1998). The situativity of knowing, learning, and research. *American Psychologist, 53,* 5–26.

Herman, J. L. (1997). Large scale assessment in support of school reform: Lessons in the search for alternative measures. *International Journal of Educational Research, 27,* 395–413.

Hickey, D. T. (2001). *Assessment, motivation, and epistemological reconciliation in a technology-supported learning environment* (Grant No. REC-0196225). National Science Foundation, Division on Research, Evaluation, & Communication to the University of Georgia.

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

Hickey, D. T. (2004). *Technology-supported multi-level assessment for improving mathematical teaching, learning, and achievement.* Research grant from the 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., 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, CA.

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.

Kellaghan, T., Madaus, G. F., & Raczak, A. (1996). *The use of external examinations to improve student motivation.*_Washington, DC: American Education Research Association.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining our schools.* Portsmouth, NH: Heinemann.

Kohn, A. (2004). Test today, privatize tomorrow: Using accountability to "Reform" public schools to death. *Phi Delta Kappan, 85*(8), 568–574.

Kolodner, J. L., Gray, J.& Fasse, B. B. (2003). Promoting transfer through case-based reasoning: Rituals and practices in *Learning by Design* classrooms. *Cognitive Science Quarterly*, 2(2).

Lagemann, E. (1999). An auspicious moment for education research? In E. Lagemann & L.S. Shulman (Eds.), *Issues in education research: Problems and possibilities*, (pp. 3–16). San Francisco: Jossey-Bass.

Lave, J. (1991). Situating learning in communities of practice. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 63–82). Washington, DC: American Psychological Association.

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

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher*, 23(9), 4–14.

Linn, R. L. (2001). A century of standardized testing: Controversies and pendulum swings. *Educational Assessment*, 7(1), 29–38.

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.*

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Minstrell, J. (2000). Student thinking and related assessment: Creating a facet assessment-based learning environment. In J. Pellegrino, L. Jones, & K. Mitchell (Eds.), *Grading the nations report card: Research from the evaluation of NAEP.* Washington, DC: National Academy Press.

Mislevy, R. J. (1993). Foundations for a new test theory. In N. Fredericksen, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 19–40). Hillsdale, NJ: Erlbaum

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2002). One the roles of task model variables in assessment design. In S. Irvine & P. Kyllonen (Eds.), *Generating items for cognitive tests: Theory and practice* (pp. 97–128). Mahwah, NJ: Erlbaum.

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. Chudowsky, N., & R. W. Glaser, R. (Eds.). Washington, DC: National Academy Press. Available online at http://www.nap.edu/catalog/10019.html

National Research Council. (2002). *Scientific research in education.* In R. J. Shavelson & L. Towne (Eds.), *Committee on Scientific Principles for Education Research.* Washington, DC: National Academy Press.

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

Natriello, G. (1987). The impact of evaluation processes on students. *Educational Psychologist, 22,* 155–175.

Ochs, E. (1996). Linguistic resources for socializing humanity. In J. J. Gumperz & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 407–437). Cambridge, UK: Cambridge University Press.

Olson, L., & Viadero, D. (2002, January 30). Law mandates scientific base for research. *Education Week* [Online]. www.edweek.org.

Pea, R. (1985). Beyond amplification: Using the computer to reorganize mental functioning. *Educational Psychologist, 20,* 167–182.

Pellegrino, J. W. (2004). Designs for research on technology and assessment: Conflicting or complementary agendas? In B. Means & G. Haertel (Eds.), *Using technology evaluation to enhance student learning* (pp. 49–56). New York: Teachers College Press.

Pellegrino, J. W., Baxter, G. P., & Glaser, R. (1999). Addressing the "Two Disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education, 24,* 307–353.

Quellmalz, E., Schank, P., Hinojosa, T., & Padilla, C. (1999). *Performance assessment links in science* (ERIC/AE Digest Series EDO-TM-99-04). ERIC Clearinghouse on Assessment and Evaluation, University of Maryland.

Resnick, L. B., & Resnick, D. P. (1992). Assessing the thinking curriculum. In B. Gifford & M. O' Conner (Eds.), *Changing assessment: Alternative views of aptitude, achievement, & instruction* (pp. 37–76).

Rogoff, B. (1998). Cognition as a collaborative process. In W. Damon, D. Kuhn, & R. Seigler (Eds.), *Handbook of child psychology* (5th ed., Vol. 2, pp. 679–744). New York: Wiley.

Ruiz-Primo, M. A., & Shavelson, R. J. (1996). Problems and issues in the use of concept maps in science assessment. *Journal of Research in Science Teaching, 33,* 569–600.

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.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18,* 145–165.

Schafer, N. J., 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.

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

Schank, R., & Ableson, R. (1977). *Scripts, plans, goals, and understanding.* Hillsdale, NJ: Erlbaum.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19,* 404–450.

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

Solono-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical, and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16–25.

Stanovich, P., & Stanovich, K. (2003) *Using research and reason in educational research: How teachers can use scientifically based research to make curricular and instructional decisions.* Washington, DC: U.S. Department of Education.

Stevens, R., & Palacio-Cayetano, J. (2003). Design and performance frameworks for constructing problem solving simulations. *Cell Biology Education, 2,* 162–179.

Stein, M., Silbert, J., & Carnine, D. (1997). *Designing effective mathematics instruction. A direct instruction approach.* New York: Prentice Hall

Stiggins, R. J. (2001). *Student involved classroom assessment* (3rd ed.). Columbus, OH: Merrill Prentice Hall.

Taasoobshirazi, G., Zuiker, S., & Hickey, D. T (2004). *Design-based implementation and evaluation: Astronomy Village: Investigating the universe. 2003–2004 implementation report.* Unpublished project report. University of Georgia Learning & Performance Support Laboratory

Torrance, H., & Pryor, J. (1998). *Investigating formative assessment: Teaching, learning, and assessment in the classroom.* Buckingham, UK: Open University Press.

Turnstall, P, & Gipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal, 22,*

U. S. Department of Education. (2002). *Strategic plan for 2002–2007.* Washington, DC: Author. Retrieved from http://www.ed.gov/about/reports/strat/plan2002-07/index.html

U. S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user-friendly guide.* Washington, DC: Author. Retrieved from http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

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

Wenk, L., Dufresne, R., Gerace, W., Leonard, W., & Mestre, J. (1997). Technology-assisted active learning in large lectures. In C. D'Avanzo & A. McNichols (Eds.), *Student-active science: Models of innovation in college science teaching* (pp.431–452) . Philadelphia: Saunders.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, & metacognition: Making science accessible for all students. *Cognition and Instruction, 16,* 3–118.

Wiggins, G. (1998). *Educative assessment.* San Francisco: Jossey-Bass.

Wiggins, G., & McTeague, J. (1999). *Understanding by design.* Washington, DC: Association for Supervision and Curriculum Development.

Wiliam, D., & Black, P. (1996). Meanings and consequences: a basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal, 22,* 537–548.

Wolfe, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17,* 31–74.

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

Zimmerman, B. J. (1993). Commentary. *Human Development, 36,* 82–86.

Zuiker, S. J., & Hickey, D. T. (2004, April). *Identities for knowing: analysis of discourse and transfer during collaborative formative feedback activities.* Presentation at the annual meeting of the American Educational Research Association, San Diego, CA.