103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 1

CHAPTER 10

Systems of Coherence and Resonance: Assessment for Education and Assessment of Education

PAUL G. LEMAHIEU AND ELIZABETH C. REILLY

The source chapters in this volume raise issues that collectively address the question of coherence within systems of assessment and accountability: what dimensions define coherence such that practical development efforts can hope to realize it; to what extent can assessments designed to address the needs of classroom practitioners serve the interests of those who work some distance from the classroom; and what design elements need to be addressed (and how should they be addressed) to maximize utility across levels of the system?

There is a lengthy history to this discussion. Nearly twenty years ago, LeMahieu and Wallace (1986) explored the conditions that best enabled assessments for what they termed *clinical* uses in contrast to those that addressed *evaluative* ones. Cole (1984) similarly explored whether assessments constructed for accountability purposes could ever be appropriate and effective for instructional uses. More recently, Black and Wiliam (1998), Shepard (2000), and Chappuis and Stiggins (2002) all examined similar questions. They explored what has variously been termed clinical versus evaluative assessment, instructional versus accountability testing, classroom versus large-scale assessment, or in its most contemporary form, assessment *for* education versus assessment *of* education. Whatever the terminology, the issue concerns the coherence and compatibility of assessments intended to inform intervention on behalf of the growth and development of students as opposed to assessments that inform judgments about the accomplishment or status of individuals, programs, schools, or systems.

While each of these researchers examined this issue from a different perspective and thus illuminated a portion of the considerations

Paul G. LeMahieu is the Director of Research and Evaluation for the National Writing Project at the University of California, Berkeley and the former state superintendent of education in Hawaii. Elizabeth C. Reilly is _____.

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 2

inherent in it, there emerges a general consensus across authors and across the years. Most have only modest hopes that the same assessment devices can serve both purposes, instead recognizing the legitimate need for each type of information and counseling that we establish systems made up of elements that address each purpose. These sentiments echo throughout the recent NRC report (Pellegrino, Chudowsky, & Glaser, 2001) in its call for coherence in terms of the integration of assessment frameworks and methods across all levels of the assessment system.

Guided by such sentiments, researchers and practitioners alike are struggling to conceive of such systems of assessment that realize genuine coherence across its various elements and, to the extent possible, integration those various elements. This means developing assessment strategies that potentially can serve multiple purposes to achieve coherence and resonance across disparate elements. Resonance occurs when separate elements of a complex system actively reinforce each other for their mutual benefit. While many of the authors' efforts are very thoughtful and successful in realizing whatever forms of coherence serve as a focus for their work, few approach being coherent in any broad sense. Simply stated, they lack a comprehensive exposition of this idea of coherence in assessment systems.

Each of the source chapters treats one or more dimensions of this idea of systemic coherence in the work that it describes. Some dimensions are addressed explicitly; many more are addressed implicitly throughout the chapters. It is possible to examine these chapters with an eye to synthesizing a much more elaborate view of coherence and, in doing so, articulate a reasonably comprehensive framework that can guide development efforts pursuing coherence and resonance. In this chapter, we will first examine each of the projects described in this volume to explore the elements of coherence that it manifests, how the elements are defined, and the manner in which the elements are addressed in the assessment program described. Following that review, we will augment those elements with others that seem necessary and organize them into a framework that can guide efforts at developing coherent and resonant systems of assessment. We will conclude with reflections on the application of the framework and advice to educational leaders and assessment practitioners seeking to use it.

Elements of Coherence

The work of Black and Wiliam, Forster and Masters, Frederiksen and White, Smithson and Porter, and Wilson and Draney addresses a

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 3

number of the elements of coherence and resonance between classroom and system assessment. Three projects point out the importance of consonance in the larger goals and purposes of assessment by looking for common elements that should make up classroom and large-scale assessment systems. Wilson and Draney (Chapter 6) frame the discussion with the principles derived from the National Research Council's report, *Knowing what students know: The science and design of educational assessment* (Pellegrino, Chudowsky, & Glaser, 2001). Coherence between classroom and large-scale assessment is essential, as is a comprehensive system—one that offers a full range of ways to assess progress. Finally, the principle of continuity provides evidence over time of individual student development and of educational programs' progress.

In their efforts to bridge the potential conceptual gap between system accountability and classroom assessment, Forster and Masters (Chapter 3) draw upon their decade of experience in Australian schools working on literacy achievement. They provide greater detail to the Wilson and Draney hallmarks of a comprehensive assessment system by presenting six components of an ideal system: 1) a progress map with a conceptual framework; 2) multiple assessment methods; 3) opportunities for professional development; 4) tools to help teachers assess student work; 5) procedures for collecting system-wide data; and 6) processes for reporting and monitoring achievement over time at the school and system levels.

Frederiksen and White (Chapter 4) add to the discussion of assessment system standards with two additional factors: directness and transparency. By directness, they refer to the tasks that students perform in assessments asserting a preference for direct assessment. Tasks must require the explicit application of the cognitive processes and knowledge that one wants to assess with the opportunity for students to demonstrate this visibly. Transparency refers to providing access to the processes used for interpreting the data so that "all participants have a shared basis for recognizing, discussing, and evaluating students' work" (p. xx). Black and Wiliam (Chapter 2) support the assertion that transparency is an essential component of assessment that must extend not only to the teachers (Wilson and Draney) through professional development, but also to the students.

In their discussion of the content that becomes the basis of assessments, Smithson and Porter (Chapter 5) begin with a detailed and extensive description of the history of instances wherein the interpretation of assessment information with reference to content coverage

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 4

has proved useful. They assert that in the end few things are quite so important as thoughtfulness about and understanding of what is being assessed. Early on, principally researchers attended to these matters. Over time policymakers and educational managers joined the researchers, and at present, the constellation of interested parties extends into offices of school districts and state capitals as well as into the classroom.

Forster and Masters present a way to describe areas of student learning—frequently called "outcomes" or "indicators"—that they term a "progress map." The use of these progress maps is the centerpiece of their notion of "developmental assessment," which becomes the principal means of summarizing and describing students' achievements and group trends over time. These progress maps present students' typical development from year to year in the form of outcomes for achievement. These maps serve both system accountability and individual classroom teachers. Wilson and Draney further explicate the concept of a progress map with descriptions of the progress variables that define the specifics of the progress maps. Weighing in on the discussion of what is assessed as an important element of coherence, Frederiksen and White emphasize measuring "significant work that students are undertaking in the classroom" (p. xx).

Three of the chapters—those by Frederiksen and White, Smithson and Porter, and Wilson and Draney—address the technical quality of assessment, namely, that matters of validity and reliability must be attended to. Smithson and Porter state that "the challenge to create valid accountability systems that draw upon teacher's professional knowledge in a manner that encourages their use of assessment frameworks and styles that are consistent with good practice" are "by far the greatest of the challenges faced" (p. xx). Frederiksen and White augment qualitative analysis by asserting the need to establish "the internal validity and external validity of the knowledge claims that are being made, based on evidence obtained in the assessment" (p. xx).

Most of the authors address in some fashion the matter of how data should be analyzed and interpreted. Wilson and Draney and Forster and Masters speak of the nature of evidence as it is generated by the assessment and that this evidence in the form of feedback is presented as quantitative data, numerical, simplified, and summarized through the analytic technology that is applied. Because the data are highly reduced, scores then are compared to other information to give them meaning, raising the issue of score referencing. Black and Wiliam enlarge the discussion with a call for more narrative data, as do

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 5

Frederiksen and White. Such narrative data would then, for the purposes of system assessment, need to be distilled and generalized in the form of trends, themes, and so on. Some form of scoring guide accompanies any classroom assessment, and according to Frederiksen and White, they must provide "a principled basis for making scoring inferences based on that evidence" [to address] how we can scaffold teachers' (and students') interpretations of students' work to ensure that resulting assessments are accurate and comparable" (p. xx).

A number of the projects address the role of students as active participants in assessment. While Frederiksen and White call this form of analysis "reflective assessment," in which students evaluate their own and other students' projects, Black and Wiliam refer to it as "self- and peer assessment." Both point out the value of student involvement in the assessment process. The benefits of this involvement are many, including the integration of assessment as an extension of the learning process rather than an add-on endeavor.

Collectively and individually the authors have identified many elements of assessment and offered examples of very practical means of addressing them to achieve coherence and resonance. We list and define a number of them here and describe the manner in which each is addressed by the authors:

- Goals and Purposes—the functions served by the assessment. The development of assessments must be guided by principles of coherence, continuity, and comprehensiveness (Wilson and Draney) with at least the six specific elements delineated (Forster and Masters).
- Locus of Authority—who defines content and standards for assessments. A locus of authority must be established, through curricular and content specialists (Forster and Masters) or through the use of curriculum standards (Wilson and Draney) or through standards documents, curriculum frameworks, textbooks, and teacher choice (Smithson and Porter).
- Content Assessed—what is measured. The definition of content and assessments can be guided by progress maps and progress variables (Forster and Masters, Wilson and Draney). Assessments should measure "significant" work that is happening in classrooms (Frederiksen and White) and measure the overlap with standards documents, curriculum frameworks, or instructional content as appropriate (Smithson and Porter).

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 6

- Nature of Evidence/Data—the type of data used to support inferences of various sorts. Evidence can be quantitative, simplified, organized, and presented in relational terms to give meaning (Forster and Masters, Frederiksen and White, Wilson and Draney). However, to maximize instructional utility, it should be expository, narrative, and descriptive (Black and Wiliam, Frederiksen and White).
- Technical Quality—the applicable hallmarks and standards of quality and fairness. Researchers should apply validity theory to both large-scale and classroom assessment (Wilson and Draney) and use it to augment qualitative interpretations that are undergirded by cognitive theory (Frederiksen and White).
- Score Reference—the comparison of derived scores with other information to imbue them with meaning. Scores can be compared with those of other students in the class (or other appropriate reference groups), with the student's individual performance over time (Black and Wiliam, Forster and Masters, Frederiksen and White, Wilson and Draney), or with identified content domains (Smithson and Porter).
- Analysis and Interpretation—procedures for deriving meaning from the evidence. Cognitive theory can bolster the use of scoring guides (Black and Wiliam, Forster and Masters, Wilson and Draney) and normative or modal response patterns (Frederiksen and White).

A Framework for Coherent and Resonant Systems

Based on the preceding review, we turn now to a more detailed explication and illustration of the dimensions of coherence and resonance between classroom and large-scale assessment design. We have synthesized the preceding discussion into a framework of elements along which coherence and resonance must be pursued (Table 1). The framework presents the seven elements listed above, augmented by nine others taken from the literature and from the authors' experiences in developing assessment systems that aspire to serve classroom practitioners as well as program managers, system leaders, and policy makers. Table 1 also offers very brief descriptions of desiderata for each element, in terms of both assessment *for* learning and assessment *of* learning. Before exploring particular entries in the table, a few general comments about the framework are in order.

First, the presentation is referenced to both assessment *for* and assessment *of* education. This is in keeping with the distinction offered

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 7

TABLE 1
Elements of Coherence and Resonance
Between Large-Scale and Classroom Assessment Design

Domain	Element	Qualities in assessment *of* education	Qualities in assessment *for* education
Conceptual			
	Goals / Purposes	Evaluation, certification, summary judgments, higher stakes; desire to establish long-term trends of student learning	Diagnosis, intervention, formative information, lower stakes; desire to make immediate changes in student learning
	Nature / Effect of motivation	Generate best possible scores; high stakes vis-à-vis rewards and sanctions; risk taking discouraged as potentially damaging to the demonstration of accomplishment	Improve teaching and increase learning; low stakes vis-à-vis rewards and sanctions; risk taking encouraged as useful to the revelation of the learner
	Epistemological frame of reference	Positivism, logical empiricism	Relativism, social constructivism
	Relationship of assessment to learners	Short term, distant, circumscribed, observer, "-etic"	Long term, involved, participatory, "-emic"
	Consequential validity / Desired impact	Distant impact on resource allocation, program-level goals, and designs; minimal impact in situ to control potential narrowing of educational experience	Immediate and profound shaping of educational experience in situ
Substantive			
	Locus of authority / Control over form and content	Macro level—external to classroom; experts in field and boards, "the public" as expressed through political structures determines content, benchmarks, and goals	Micro level—involving teachers and students in classroom; teachers determine content, benchmarks, and goals often using state or national standards
	Content assessed	Knowledge, skills, behaviors based on global, industrial, or societal learning outcomes external to classroom; emphasis on breadth	Knowledge, skills, behaviors based on local, individual, class, or course learning outcomes; emphasis on depth
	Methods / Instruments	Tests, scales, surveys generally added to class work; efficiency and economy highly prized	Observations, interviews, personal documents derived from class work

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 8

TABLE 1 (*Continued*)
Elements of Coherence and Resonance
Between Large-Scale and Classroom Assessment Design

Domain	Element	Qualities in assessment *of* education	Qualities in assessment *for* education
Technical			
	Nature of evidence / Data	Reduced, often quantitative, standardized, emphasis on comparability and aggregation	Narrative, extensive, "thick," "rich," potentially idiosyncratic
	Technical quality	Usually privileges reliability as well as content and concurrent validity	Usually privileges consequential and hermeneutical notions of validity
	Score reference	Normative; standards referenced	Standards referenced; criterion referenced; ipsitive
	Unit of analysis / Sample size	Groups of individuals, relatively large sample	Individual, relatively small sample
Procedural / Logistical			
	Data generation plan	Structured, predetermined, standardized, detailed plan of operation; replication and generalizability desired	Evolving, flexible, authentic conditions desired; replication and generalizability not necessary
	Timing / Frequency of data collection	Infrequent, periodic, or even ad hoc; often collected and stored over long time	Frequent even ongoing; often collected and stored for short time
	Timeliness of feedback / Data analysis	Delayed	Immediate
	Primary audiences and users	External: policymakers, education leaders and managers, the public	Internal: teachers, students, and parents

by Black and Wiliam (1998), in which assessment intended to inform decisions about educational opportunities offered to students for the purpose of strengthening their instruction and learning is referred to as assessment *for* learning, and assessment that informs decisions about the accomplishments of programs, schools, and systems is referred to as assessment *of* learning. This distinction roughly parallels that between classroom and large-scale (or clinical and evaluative, or instructional and accountability) testing.

Second, the depictions of assessment *for* and assessment *of* learning that are offered for each element are admittedly exaggerated. Few

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 9

assessments fully realize these extreme portrayals on many of the elements and none exemplifies them for all sixteen of the elements. The framework is best understood as a template for ideal practice. Any single assessment will exhibit a profile consisting of its positions on each of the continua, positions that will be closer to the assessment for education ideal for some elements and farther away from it for others. It is in the analysis of this multidimensional profile that one can judge the suitability of an assessment for one or the other purpose.

The framework presents many elements in terms by which the attributes of assessments *for* learning and *of* learning can be characterized. They are organized into four broad domains: those pertaining to the conceptual framing of the assessments; those describing its form and substance; those qualifying its technical attributes; and those concerned with its procedural and logistical administration. Within each domain exists a number of particular elements. The descriptions of the qualities of each element in terms of assessment *for* and *of* learning are not intended to be exhaustive or exceptionally detailed. Rather, they are framed so as to highlight distinctions between the two types of assessment as they relate to each element.

Upon first examination, the characterizations of ideal qualities with respect to many of the elements seem incompatible. For example, assessment *for* learning calls for very frequent (even ongoing) measurements, while assessment *of* learning requires much less frequent measures. Similarly, the timeliness of feedback must be immediate (or as close to it as is possible) in the case of assessment *for* learning and can typically be much more delayed in assessments *of* learning.

The application of the framework to an existing assessment system (a type of system with which most are familiar) is very illuminative. With this in mind, we turn to the typical state-implemented large-scale assessment system used for evaluative and accountability purposes. An exhaustive and extremely detailed characterization, however, is beyond the scope of this chapter (and probably not interesting or helpful apart from taking up a close examination of a particular state's program). While there are certainly differences from state to state, it is nonetheless possible to characterize such assessments generally and for the most part in a fair manner. Here we will describe the typical state assessment system in terms of the elements presented in the framework. However, for illustrative purposes we will examine only selected elements from each of the domains (i.e., Conceptual, Substantive, Technical, and Procedural/Logistical).

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 10

In the typical state assessment system, the element *relationship of assessment to learner* differs between classroom and large-scale assessment. Large-scale assessments tend to minimize this relationship in order to lessen the amount of instructional time given over to the assessment, or to maintain standard testing conditions across students. On the other hand, classroom assessment perspectives permit, and in the ideal encourage, a long-term, involved, and in some instances even participatory (in terms of defining the assessment itself) relationship.

The *consequential validity/desired impact* of the assessment is another element along which interesting differences arise. Both assessment types may lay legitimate claim to improving education, as each provides information that can conceivably be used for its betterment, whether directly or indirectly. Large-scale state assessment does so through more distant mechanisms: the identification of needs (and the required or desired responses to those needs), the allocation of resources, and most recently, the application of accountability influences, so as to increase and focus effort with the goal of improving performance. The consequential impact of classroom assessment is (to the teacher and learner at least) much more immediate and much more profound. It shapes the choice of curricular and instructional approaches; it contributes to the grouping and regrouping of students for instruction; and it directly influences the learning experience.

A third element in the Conceptual domain is *nature/effect of motivation.* Typically in large-scale assessments the test taker is singularly motivated to demonstrate accomplishment by generating the highest scores. This is certainly true as the political and most recently even material salience of those scores increases. Classroom assessment, on the other hand, ideally encourages risk taking and with it the occasional misstep, and it values the demonstration of current and future needs alongside the demonstration of accomplishment.

Additional similarities and differences exist across the other domains as well. For example, in the Substantive domain there are generally differences between these two types of assessment with respect to *locus of authority/control over form and content*. A state's large-scale assessment *of* learning typically vests such authority external to the classroom to maintain common standards across settings and to ensure the validity and credibility of accountability processes. By contrast, classroom assessment *for* learning places authority with teachers and students in the classroom, with interesting (some would say beneficial, some not) social, political, and educational consequences. With respect to *nature of evidence/data*, in the Technical domain, there are

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 11

typically differences as well. Large-scale assessments have an obligation to serve individuals and groups who work at some distance from the classroom. Therefore, they typically prize highly reduced information. Classroom assessments designed to inform and drive instruction seem to do so best when they deliver thick, rich, descriptive narratives of performance, needs, and means of addressing them.

Reflections and Considerations

Even just this brief application of the framework to a typical state assessment system in contrast to a typical classroom assessment reveals much about the framework itself and the assessments that it describes. First, the very idea of coherence is a complex matter. More than attending to similarities of content, coherence is best understood as requiring critical analysis and effort across a number of elements. Second, while there are clearly differences between large-scale and classroom assessment across these many elements, many are by no means intractable. Innovative thinking coupled with hard work in expanding the science of assessment development opens up the possibility of coherence (or, beyond coherence, resonance in the form of mutual support and benefit) far surpassing anything that we have achieved to date.

Thus the framework raises the question of whether assessments can be designed, constructed, and implemented in ways that serve both ends. While this is an important question to confront, the answers are not as unambiguous as first examination of the framework might suggest. Three factors open up possibilities for coherence and genuine resonance, and within them, mutual benefit across assessment types.

First, the characterizations of the elements must be regarded as a mixture of descriptive and prescriptive statements. Thus, while it is clearly a prescription that assessment *for* learning involves immediate feedback (as much of the research on feedback in the learning process supports this view), it is by no means a prescription that there be delays in the reporting of results in accountability systems. This is rather a descriptive statement of typical practice. Subject only to technological and practical exigencies, assessments *of* learning could serve classroom purposes better if they could be implemented in ways that provided feedback much more quickly. Here the framework challenges assessment developers to use innovations in practice that would greatly enhance the utility of the assessments. Not all of the elements are as clearly or easily determined to be prescriptive or descriptive. Only extensive efforts at innovative development and attendant

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 12

research and reflection (such as those described in the source chapters of this volume) will reveal the limits and possibilities.

A second factor enhancing the potential for coherence and resonance across these elements is that they are by their nature interdependent and to some extent compensatory. Thus, while the statements about the consequential validity and desired impacts of accountability assessments may lead to problematic and troubling consequences in the classroom (as so many authors have pointed out), the extent to which they necessarily do so depends upon the profile of the assessment across many of the other elements. To the extent that an assessment *of* learning assesses domains of genuine interest in the classroom or does so using item types that are reflective of desired practice there, it may be less troubling to see somewhat more direct impact upon classroom practice. The only way to speculate with confidence (as guidance for development efforts) is through analyses guided by the framework; the only way to know with certainty is through empirical investigation.

A third factor revealed through close analysis of these elements is that thoughtful and clever design can enable an assessment intended primarily for one purpose to provide evidence serving the other. The timing and frequency of assessments discussed above is a good example. While the assessments that serve classroom decisions and practice should be administered frequently (much more so than those for accountability), it is possible to design for classroom purposes with respect to this element and occasionally cull evidence for accountability purposes as appropriate.

Similarly, ample research (including several of the source chapters in this volume) demonstrates the nature of evidence that serves each purpose best. Assessment *for* learning requires descriptive, narrative feedback—what some have termed thick, rich, or extensive. By contrast, most of the evidence found useful by those who work at a great distance from the classroom is highly reduced. Clever design might permit generation of evidence useful in the classroom before subsequent analyses reduce and summarize that evidence in ways appropriate to accountability systems. Similar treatment could well address nearly all of the elements listed in the Technical and Procedural/Logistical domains of the framework.

It is interesting to note that in many cases there is a directionality or asymmetry to the flow or processing of evidence. In every regard, assessment *for* learning requires evidence that is more complex, detailed, persistently gathered, and timely. This framework and the

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 13

authors' personal experience in assessment development suggest that to the extent that there is interest in expanding the purposes to which assessment information is appropriately put, synchronicity is most likely found in developing assessments to serve classroom purposes first, and then processing the evidence in ways that make it appropriate to and useful for accountability purposes. (See, for example, LeMahieu, Eresh, & Wallace, 1992; LeMahieu, Gitomer, & Eresh, 1995; LeMahieu & Eresh, 1996 for another example of assessment development guided by these principles to serve expanded purposes.)

The initiatives described in the source chapters of this volume explore the extent to which assessments can be pushed to serve multiple purposes. From these descriptions, we have distilled a framework that enumerates and organizes the elements to which efforts at coherence must attend. These initiatives begin to illustrate in practical terms the ways in which such coherence can be pursued. There is much more that can be done. Each case that seeks to expand the utility of classroom assessment in service of accountability does so with the well-intended goal of rendering accountability contingent upon assessments that are coherent with classroom goals and practices. However, a certain cautionary tone in the authors' writing betrays a reluctance to commit to the position that one assessment can serve both purposes wholly.

It would seem that classroom assessments pressed into the service of accountability (especially when made politically charged by systems that involve rewards and sanctions) might be rendered useless as assessments for the classroom context. Many have written of the potential for content choices that distort the curriculum, particularly as the political salience of the assessed content is heightened. At the very least, the accountability motives would seem to militate against the risk taking necessary to reveal completely the students' learning needs.

In the face of such potentially disturbing consequences, it would seem that the best accountability assessment is one that describes the status or accomplishment of its focus while not influencing what is measured at all. However, Heisenberg makes clear that this is quite impossible. Moreover, policymakers who invoke accountability as an essential element of reform rather hope that it will exert such influence.

One outcome of the application of the framework presented here is to expand the range of utility for both kinds of assessment. It instructs and guides the design of assessments that remain true to their first purpose while being expansive along the sixteen elements in ways

103-2 Ch10 LeMahieu & Reilly.qxd 1/15/2004 1:22 PM Page 14

that define those assessments as more appropriate to certain purposes. The framework also challenges our development of assessments to move beyond mere coherence, defined as noninterference, and to achieve a resonance in complex systems in which the parts are mutually supportive and beneficial.

References

Black, P.J., & Wiliam, D. (1998). *Inside the black box: Raising standards through classroom assessment*. London: King's College School of Education.

Chappuis, S., & Stiggins, R. (2002). Classroom assessment for learning. *Educational Leadership*, *60*(1), 40-43.

Cole, N.S. (1984). Testing and the "crisis" in education. *Educational Measurement: Issues and Practice*, *3*(3), 4-8.

LeMahieu, P.G., & Wallace, R.C., Jr. (1986). Up against the wall: Psychometrics meets praxis. *Educational Measurement*, *5*(1), 12-16.

LeMahieu, P.G., Eresh, J.T., & Wallace, Jr., R.C. (1992). Using student portfolios for public accounting. *The School Administrator: Journal of the American Association of School Administrators*, *49*(11), 8-15.

LeMahieu, P.G., Gitomer, D.A., & Eresh, J.T. (1995). Portfolios in large-scale assessment: Difficult but not impossible. *Educational Measurement: Issues and Practice*, *13*(3), 11-28.

LeMahieu, P.G., & Eresh, J.T. (1996). Comprehensiveness, coherence and capacity in school district assessment systems. In D.P. Wolf and J.B. Baron (Eds.). *Performance based student assessment: Challenges and possibilities. The Ninety-fifth Yearbook of the National Society for the Study of Education*, Part I (pp. 125-142). Chicago: National Society for the Study of Education.

Pellegrino, J.W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

Shepard, L. (2000). The role of assessment in a learning culture. *Educational Researcher*, *29*(4), 4-11.