ASSESSING FOR LEARNING: SOME DIMENSIONS UNDERLYING NEW APPROACHES TO EDUCATIONAL ASSESSMENT

John Biggs
University of Hong Kong

Abstract

The theory and practice of assessing learning are currently undergoing a paradigm shift. The critical realization in producing this change is that educational considerations should drive testing, not psychometric or political ones. Three dimensions interact to yield different modes of assessment, including different kinds of performance assessment: the measurement vs. the standards model of testing, quantitative and qualitative assumptions as to the nature of what is learned, and whether the learning and testing is situated or decontextualized. The modes of assessment so generated are suited for different educational aims, but the most appropriate modes are under-represented in current practice, quantitative and decontextualized modes being greatly over-represented, resulting in backwash often deleterious to teaching and learning. Conceptual and structural difficulties in implementing qualitative and situated modes of assessment are discussed.

Alberta Journal of Educational Research, 41, 1 - 18, 1995

Address for correspondence: Professor J.B. Biggs,
Education Department,
University of Hong Kong,
Pokfulam Road,
Fax: (852) 8585649 Hong Kong.

ASSESSING FOR LEARNING: SOME DIMENSIONS UNDERLYING NEW APPROACHES TO EDUCATIONAL ASSESSMENT

John Biggs
University of Hong Kong

Assumptions Underlying Assessing and Learning

The measurement model

For many years, perhaps the greater part of this century, the assessment of academic learning proceeded with little change and with little perceived need for change. Test developers, psychometricians, and teacher educators accepted the *measurement* model of assessment. This model is based on trait theory, which requires that test item scores are sufficiently consistent with each other so they may be assumed to lie along a single dimension, and that testees may be ordered along the dimension so represented (Taylor, 1994). The technologies of test construction, item selection, and establishing reliability and validity followed from these basic assumptions. That technology became very sophisticated, and powerful in its appropriate applications. Classroom teachers, for their part, espoused the measurement theory taught them in their teacher education , but their theory-in-use was an uneasy compromise between that and established practice (but see Cizek, 1993); when they do try to put their espoused theory directly into practice they usually get it wrong (Marso & Pigge, 1991).

The problem is not teacher incompetence, but the fact that the "technology of assessment that grew out of test theory...lacked a basis in psychological theory" (Wilson & Kirby, 1994: 107); even more to the point, it lacked a basis in educational theory and the knowledge base of teaching (Haertel, 1991). Traditional test theory became, with changing educational philosophies, decreasingly appropriate to the majority of classroom testing occasions. Specifically, the model assumes the stability of the dimension being tested, which is very appropriate when the task is to discriminate between individual performances and to predict future performance, but is inappropriate to assess individual or group outcomes in relation to a particular curriculum because the intervention of *teaching* assumes change. As social expectations became increasingly that most students should complete grades K through 12, acquiring basic skills, competencies, and certain declarative knowledge on the way, selecting students for different levels of schooling ceased to be a major issue, but ensuring that students attained these standards acceptably (however one defines that in the event) was a definite issue of public concern. In an age of electronic learning, assessment practices were being driven by steam.

The standards model

Models of assessment based on outcomes reaching a predetermined standard have also been around for many years—for example, the notion of a "First Class Honours" in the British University system—but these ideas were very subjective and they specified no technology. The criterion-referenced testing (CRT) movement was probably the first systematic attempt to formulate and to enact a set of assumptions critically different from those underlying the measurement model, in that instructional and assessment

strategy were specifically linked (Bloom, Hasting, & Madaus, 1971; Popham & Husek, 1969). CRT is an example of what Taylor (1994) calls the *standards* model of assessment, which is based on such assumptions as: public standards can be set; they can be reached by most students, albeit by different kinds of performance; and fair and consistent judgments are possible to determine whether the standards have been met or not. These assumptions about assessment are isomorphic to those underlying learning and instruction.

The traditional CRT model fell short, however, on the *nature* of the standards that were set, which in some important respects were no improvement upon those underlying the measurement model itself. While CRT was an improvement in terms of the link between assessment and instruction, the links between these and the question of what is learned remained unchanged. CRT had as limited a view of learning as had the measurement model.

Two basic conceptions of the nature of learning exist in our educational thinking, quantitative and qualitative (Cole, 1990; Marton, dall'Alba, & Beaty, in press). How we view learning will determine how we go about learning, teaching, and, to the present point, assessment.

The quantitative tradition

The quantitative tradition has the longest history in educational thinking, stemming from the positivist tradition in the social sciences (Moss, 1992). Learning is conceived as acquiring "specific discrete skills described as precise well-delimited behaviors" (Cole, 1990:2). These contents of learning are treated as discrete quanta of declarative or of procedural knowledge; as far as assessment is concerned, any one quantum is treated as functionally independent of any other. In this view, the curriculum becomes in effect a list of discrete units: facts, skills, competencies, behavioural objectives, performance indicators, and the like, and assessment a matter of how many of these have been attained.

Teaching is conceived as transmitting knowledge from teacher to learner, and many delivery and assessment systems are based on the transmission model, up to college and university level (Trigwell, Prosser, & Taylor, 1994). The teacher's task is to know the subject and expound it clearly, the learner's to receive accurately. In assessment practice, the contents of knowledge are seen as learned in binary units (correct/incorrect), the correct units being summed to give an aggregate score that yields an index of competence in what is learned.

Multiple choice tests enact this clearly; competence is represented as a total score of all items correct, any one item being "worth" the same as any other. Lohman (1993) cites an example of a multiple-choice test given to fifth-grade children when the 200th anniversary of the U.S. Constitution was being celebrated. The only item on the test referring to Thomas Jefferson was: "Who was the signer of the constitution who had six children?" The problem is that this tells the child that every idea in the test is equally important. Lohman recounts that a year later he asked a child in this class what she remembered of Thomas Jefferson: the number of his children, but nothing of his role in the Constitution. Such testing tells students that:

> There is no need to separate main ideas from details; all are worth one point. And there is no need to assemble these ideas into a coherent summary or to integrate them with anything else because that is not required. (Lohman, 1993: 19).

However, particularly in large classes where marking schemes are used, essays are frequently treated in the same manner, a mark being given as each "correct" or "acceptable" point is made, with possibly bonus points for argument, or style. I am not saying that individual items in an MC test may not be clever, demanding, or substantively significant, or that the essay question is unsuccessful in eliciting high cognitiv/e level engagement, but that the *treatment* of the item scores and test marks assumes their mutual equivalence, independence, and additivity. Students know this, and are strategic in exploiting it (Biggs, 1973; Crooks, 1988); in timed examinations, for instance, in view of the law of diminishing returns with respect to mark allocation (time spent on the first half of an essay almost always nets more marks than the same time spent on the second half), attempting all five, say, questions but finishing none will capture more marks than writing a properly structured answer to only four. It is interesting to note that the national model in the U.S. for setting both objective and essay questions, the Bloom taxonomy, "has no category for the organization of factual knowledge" (Lohman, 1993: 22).

Nevertheless, whether the backwash effects of MC and other testing formats in the quantitative tradition are seen as deleterious or beneficial for instruction is a much debated question, the answer no doubt depending on whether one holds a quantitative view of learning or not. There have been several warnings of the harmful influences, ranging from the popular (Hoffman, 1962) to the scholarly (Frederiksen, 1984), while psychometricians who hold avowed quantitative assumptions about the nature of learning see deliberate and systematic teaching to the test as sound teaching (Shepard, 1991), elevated by Popham (1987) to the strategy of "measurement-driven instruction". It will be noted that quantitative assumptions underly both norm-referenced testing in the traditional measurement model and criterion-referenced testing in the Popham and mastery models. The issue is thus not simply the measurement versus the standards models (Taylor, 1994), but a question of one's assumptions about the nature of the learning to be assessed.

The qualitative tradition

The qualitative tradition has its roots in nineteenth century phenomenology, and later in Gestalt psychology, but it is only quite recently that it has got as far as offering an alternative paradigm as far as educational decision-making on a wide front is concerned. The underlying theory of learning is *constructivism*, a family of theories rather than any one, according to which students are assumed to learn cumulatively, actively interpreting and incorporating new material with what they already know. Different theories variously emphasize the individual, social, cognitive, saccadic, contextual or emergent natures of learning, but all agree on an active learner seeking meaning by constructing knowledge rather than by receiving and storing knowledge. Post-structuralists emphasize the social construction of knowledge (Delandsheere & Petrovsky, 1994), while my own view, and that underlying the present paper, stems ultimately from cognitive psychology (Biggs & Moore, 1993). In this latter view, understanding changes progressively as people learn, with qualitative changes taking place in the nature both of what is learned, and how it is structured. Understanding of a topic thus evolves

cumulatively over the long haul, having "horizontal" interconnections with other topics and subjects, and "vertical" interconnections with previous and subsequent learnings in the same topic.

As the contents of learning are meanings, the curriculum question is to decide what meanings or levels of understanding are "reasonable" at the stage of learning in question. The teacher's task is not then to transmit correct understandings, but to help students construct understandings that are progressively more mature and congruent with accepted thinking, recognizing that in many subject areas students' everyday experiences have helped them to construct alternative ways of construing their world. Teaching techniques may be expository at times, but essentially they will involve more effective ways of eliciting constructive activity on the part of the student, such as the deliberate and explicit use of the relevant knowledge base, peer and student-teacher interaction, a motivating context, and much student activity, both reflective or self-directed as well as task-directed (Biggs & Moore, 1993).

Whereas the logic of assessment from a quantitative point of view implies aggregating units of learning taken cross-sectionally with respect to time, that from the qualitative tradition implies charting longitudinal growth over time, from relative ignorance to relative competence: establishing the limits of "relative" is of course a major curriculum question. The outcomes of learning become the constructions the learner has made at any given point in the process. If that growth in competence can be described in recognizable stages then so much the better, because these stages can then become assessment targets (Biggs & Collis, 1989; Scarino, Clark, & Brownell, 1994).

Assessment within the qualitative framework may be of two basic kinds:

1. *developmental*, the purpose of which is to discover where students are in the development of understanding or competence within the domain of the concept or skill in question, the focus here being pure or discipline based knowledge;

2. *ecological*, the purpose of which is to discover if students can carry out tasks that are "worthwhile, significant, and meaningful" (Archbald & Newman, 1988:1), the focus here being on applications and problem solving.

As far as issues of assessment are concerned, there is some similarity between quantitative/qualitative distinction, and Lohman's distinction between crystallized abilities as being involved in teaching and assessing for near transfer in familiar situations, and likewise for fluid abilities for far transfer in unfamiliar situations; the first involve lower order, and the second, higher order cognitive skills. Education should involve both, but common assessment tasks evoke predominantly the former (Lohman, 1993).

Situated and decontextualized assessment

Ecological assessment thus becomes the qualitative assessment of applied procedural knowledge, and as such is closely related to the more general movement referred to as "authentic" assessment (Newman & Archbald, 1992; Wiggins, 1989), which insists that the context of testing should reflect the goals of learning, in so far as they require students to think, decide, and act in the real world (Archbald & Newman, 1988).

However, the term "authentic" grabs the moral high ground rather, and "performance" assessment is now more usual for this mode of assessment (Moss, 1992), being neutral with respect to rectitude, while suggesting that the test items should require some kind of active demonstration of the knowledge in question rather than a propositional account of it. Some action is required in a realistic setting, involving enactment of a skill or problem solving. Conventional pencil-and-paper tests evidently do not meet this requirement.

Performance assessment (PA) is closely related to the concept of situated cognition. Brown, Collins, and Duguid (1989) suggest that the only valid or powerful form of learning is that which takes place in "situated" contexts. Schools, on the other hand, assume "a separation between knowing and doing, treating knowledge as an integral, self-sufficient substance, theoretically independent of the situations in which it is learned and used. The primary concern of schools often seems to be the transfer of this substance..." (Brown, Collins & Duguid, 1989: 32). These authors do not consider the declarative knowledge taught in schools to be "robust", so that in continuing to focus on the transfer of such knowledge, the school culture becomes "inauthentic", providing students with *ersatz* activities. Schools should instead provide students with a context and activities that lead to the construction of knowledge as it used (*op. cit.);* the notion of performance assessment thus appears integral to this position.

While there is no doubt that context-based learning is very much more powerful than learning disembedded content, there is a place in school for learning decontextualized content; indeed, it could be argued that that is what schools are mainly for (Biggs, 1992a). If knowing and doing were inseparable, there could be a problem in accounting for civilization; to know only through doing virtually requires each generation to reinvent the wheel. Perhaps then we are talking about two different things: the status of different *kinds* of knowledge, and the most efficient *means of learning* any kind of knowledge.

Whatever means of teaching we adopt—inductive, problem-based, hands-on, on the one hand versus expository on the other—there is much knowledge, school-delivered, that is legitimately propositional or declarative, and which needs to be assessed as such, as well as being assessed in its real world applications. Thus, while learning takes place most *easily* in situated contexts, schools exist precisely to help students learn decontextualized second-order symbol systems, and the declarative and procedural knowledge encoded by them. Learning these systems and contents is unlikely to occur in children's direct experience of the world, and it does not come easily, possibly for biological reasons (Biggs & Moore, 1993), but such knowledge needs to be learned, and the quality of that learning needs to be assessed.

In short, then, there is a place for decontextualized learning and assessment, just as there is for situated. So far, the assessment of decontextualized learning has been in practice greatly over-emphasized, quite out of proportion to its place in the regular curriculum, but it would be equally unbalanced to entertain only situated learning and assessment contexts.

Let me now refer to a model that provides a useful framework for both modes of assessment: for charting the course of developmental assessment, and for providing a

generalizable language for discussing student performance in at least some performance assessment tasks (I would prefer my own term "ecological", but "authentic" has only just been deposed and further semantic quibbling is undesirable at this stage).

A Generalized Model of Qualitative Assessment

In the developmental model of assessment, it is first necessary to chart the course of development of a concept or principle, so that the stages of development can be defined, and the level at which a student is currently thinking determined. We thus need to describe what the learning will be like at any particular stage in its growth. This may be done on a topic by topic basis, as has been been for some topics by Marton and his coworkers using the techniques of phenomenography (Marton, 1988; Ramsden, 1988), which involves probing interviews that usually reveal layers of understanding of the target concepts: a hierachy of conceptions that can be used to form assessment targets. A more general model would require us to define increasingly higher quality in terms of such as aspects as: increasing complexity of structure, abstractness, economy or elegance of processing, originality of the response, and so on. Quality involves many different aspects, according to the task in question, but one aspect that is common to most tasks is the *structural complexity* of the learning outcome. Basically, there are two aspects to structural complexity: the amount of detail in the student's response (the quantitative aspect), and how well put together that detail is (the qualitative aspect). Both aspects are important, and may be classified by the SOLO Taxonomy (Biggs & Collis, 1982; 1989).

"SOLO", which stands for **S**tructure of the **O**bserved **L**earning **O**utcome, provides a systematic way of describing how a learner's performance grows in complexity when mastering many tasks, particularly the sort of tasks undertaken in school. A general sequence in the growth of the structural complexity of many concepts and skills is postulated, and that sequence may be used to guide the formulation of specific targets or the assessment of specific outcomes.

1. The task is not attacked appropriately; the student hasn't really understood the point and uses too simple a way of going about it (prestructural).

2. One (unistructural), then several (multistructural), aspects of the task are picked up and used, but are treated independently and additively. Assessment of this level is primarily quantitative.

3. These aspects then become integrated into a coherent whole (relational); this level is what is normally meant by an adequate understanding of the topic. Assessment of this level becomes qualitative, if it is to pick up its nature.

4. The previous integrated whole may be conceptualized at a higher level of abstraction and generalized to a new topic or area; this too requires qualitative assessment.

These levels, and the general structures incorporated within each, provide a basis for assessing the quality of particular learning episodes. It should be noted that while levels (1) and (2) may be assessed quantitatively and additively, (3) and (4) require an interpretive or hermeneutic approach (Moss, 1994); it is not the components themselves that determine the quality of the outcome but the whole that their interaction and

integration defines.

The SOLO taxonomy may be used in two kinds of assessment format:

1 *Assessing open-end outcomes* is a procedure that is straightforward and moderately well documented (Biggs & Collis, 1982). Its application to marking assignments in a standard letter-grade system is given in Biggs (1992b). While SOLO may not exactly provide a common currency across tasks, it does provide a common way of thinking about performance in quite different tasks, which is why it may be particularly useful in assessing, for example, students' portfolios. The justification of each performance the student provides in a portfolio, and the pattern formed by the selections as a whole, will exemplify a SOLO structure, which will say something about the way that student thinks about the course: as several unrelated tasks or performances, or as different key tasks that reveal an integrated way of conceptualizing the course, or even inventive applications or generalizations that go beyond the course itself.

2. *Assessing in an objective-type format*, the "ordered outcome" format (Masters, 1987), is less well documented, and needs some discussion here.
The ordered outcome format looks like a multiple choice item without the choice: the subitems are ordered in a hierarchy of competence, and all require a response, each indicating a particular level in the competence hierarchy. Masters (1987) used a phenomenographic hierachy in his prototype of the format (see above; Marton, 1988), but the disadvantage is that phenomenography is highly content specific and may yield any number of levels, which makes it complicated to use in practice. Using SOLO, the following criteria emerge for subitems addressing the stem topic at each SOLO level:

1. Unistructural: Contains one obvious piece of information coming directly from the stem.
2. Multistructural: Requires using two or more discrete and separate pieces of information contained in the stem.
3. Relational: Ues two or more pieces of information each directly related to an integrated understanding of the information in the stem.
4. Etended Abstract: Requires use of an abstract general principle or hypothesis which can be derived from, or suggested by, the information in the stem.

To illustrate, an ordered-outcome mathematics test was given to several hundred Year 7 students in each of two Hong Kong schools (Biggs, Lam, Balla & Ki, 1988). The content of the test is unremarkable in itself, but the responses to one item are revealing in the present context (see Figure 1):

-- Figure 1 goes here—

The two schools perform similarly up to multistructural level, but they diverge sharply thereafter, 48% of School B students obtaining correct on the extended abstract subitem: 48% correct compared to 6% of School A; the differences between the students in Schools A and B, whatever their genesis might be, are reflected only in the most complex cognitive processes. A conventional test, comprising an aggregate of mixed items scored correct or incorrect, would be unlikely to pick up this qualitative difference

in the students' mathematical thinking. Further, there is little reason why a teacher of grade 7 would normally think to test beyond a multistructural level. Here, the ordered outcome format forces the teacher to "think upwards" when designing the test, and hopefully likewise the student when responding.

In pointing out the two traditions in educational thinking, quantitative and qualitative, Cole (1990) points out that each has had a useful history and a continuing presence, but as that each appears incompatible, the "public understanding of education is hurt by allowing these two unconnected conversations about educational achievement to continue separately" (*op. cit.*: 5), so that an integrating framework is necessary to allow each to persist where it is effective and appropriate. SOLO appears to provide such an integration: in the early learning of many topics, quantitative aggregation is an appropriate or convenient way of assessing, but for higher order, applied, and critical thinking (relational and extended abstract) qualitative assessment is more appropriate. The important thing is that modes of assessment appropriate to lower levels do not preclude higher levels, or suggest to students that they can meet requirements by substituting lower for higher levels, which is the usual problem with backwash.

Backwash: Problem or Solution?

Let me then return to the question of backwash: the notion that testing drives not only the curriculum, but teaching methods and students' approaches to learning, usually adversely (Crooks, 1988; Elton & Laurillard, 1979; Fredericksen, 1984; Frederiksen & Collins, 1989).

These observations on the effects of backwash have been largely of the traditional measurement model framework, or in CRT conceived quantitatively. In the last case, backwash is explicit and actively encouraged in so far as the test deliberately becomes the target for teaching, as in measurement-driven instruction (MDI) (Popham, 1987). An important value question is raised here, as there is evidence that the success of such a strategy depends on how students typically go about learning: those who typically focus on and reproduce detail (a "surface" approach to learning) like the strategy and do well, but those who adopt a more academic or "deep" approach, originally better than surface learners, become frustrated and do progressively worse (Lai & Biggs, 1994).

The problem here is simply that the target of learning defined in the quantitative framework is of a low cognitive level, covering only the first two levels in the SOLO taxonomy. Would targets of higher cognitive levels promote backwash beneficial to instruction and encourage students to adopt "deeper" approaches to learning? There is some evidence suggesting this to be so. Tang (1991), for example, showed that different modes of assessment, final (short-answer) examination and a single topic assignment, elicited quite distinct assessment preparation strategies, the assignment generally producing higher cognitive level strategies based upon wide reading, collaborative learning, and problem solving, but over-riding the question of the actual mode of assessment was the student's *perception* of what was required for optimal results. Further, students needed to have the procedural knowledge necessary to enact those perceived requirements; extended writing was a novel task to many of Tang's students, and lack of the appropriate skills prevented them from realizing what they could see as required.

Higher cognitive level targets need therefore to be perceived as requiring high

cognitive level preparation strategies that are within the student's repertoire. Ordered outcome testing should send the message that what is important is to think in *increasingly complex* ways about a topic, not to obtain a certain number of correct items. The gap between a student's best response and the highest response in the hierarchy tells both teacher and student what still remains to be learned. Wong (1994) set parallel forms of a grade 11 math test using the traditional format and quantitative scoring, and the ordered-outcome format; he then interviewed students while they solved the different item-types. The difference he describes as that between "novices" (solving correctly but algorithmically) and "experts" (solving economically and originally from first principles), yet it was the same student who was both novice and expert, the variable being the test item type. It is important that further such studies be carried out, as this would provide much needed empirical as opposed to rhetorical justification for qualitative approaches to assessment.

Dimensions of Assessment

Performance/authentic assessment, then, is but one aspect of the development from traditional testing to the current situation with respect to alternative modes of assessment. To put this in perspective, at least three dimensions are involved in this evolution:

1. The *function* of testing: is it to rank individuals along some assumed trait, as in Taylor's measurement model, or to refer an individual's performance to a standard?

2. The nature of what it is that is to be assessed: a unitized fragment of performance or the growth of understanding?

3. The context in which the test item is placed: is it embedded in a context isomorphic to that in which the knowledge is or will be used in everyday life, or is it abstract and decontextualized?

Table 1 puts these points together.

-- Table 1 goes here—

The qualitative-quantitative dimension establishes the nature of what is to be assessed and how it may be reported. The measurement and standards models may operate within the quantitative, yielding NRT and CRT as traditionally implemented, but only the standards model may operate within the qualitative dimension, because NRT requires unidimensionality, whereas assessing an outcome qualitatively is a hermeneutic not a dimensional or measurement-based process (Moss, 1994). The original "authentic" debate raised the issue of whether assessment might best be situated or decontextualized. Decontextualized assessment would include the usual method of pencil-and-paper testing, which may be construed within either a quantitative framework, as is traditionally has been the case, or a qualitative framework, as is the case with most SOLO testing to date, and with ordered-outcome testing. We then need to distinguish between testing the student's developing understanding of a concept, particularly but not essentially of declarative knowledge, and the student's ability to involve that knowledge in a task that has ecological validity with respect to the learning goals. The issue in the last case is not so much what kind of understanding students have of the content, but

whether the taught content can empower decision-making in a real context (Maguire, 1990; Masters & Hill, 1988; Wiggins, 1989).

We are, then, left with six cells indicating modes and contexts of assessment.

1. *Quantitative-Measurement-Decontextualized*: Traditional NRT, which now has a very sophisticated technology and the overwhelming approval of the measurement establishment and of most administrators. Best used for selection and other individual or group comparative purposes, relatively curriculum-free. The backwash generated is almost certainly the most deleterious to teaching and learning of all six modes.

2. *Quantitative-Standards-Decontextualized*: "70s" style CRT, mastery learning. Linked very closely to curriculum and instructional strategy. Best used for testing basic or core skills, behavioral skills. The quantitative framework however limits its application to lower cognitive levels. Backwash suits students preferring a surface approach to learning, but counter-productive for students preferring a deep approach (Lai & Biggs, 1994). A model that in many senses straddles (1) and (2) is Item Response Theory (IRT) or Rasch Model (Wright & Stone, 1979), which is quantitatively framed, and uses a version of trait theory, but the scoring does not depend on how other students perform, as in NRT, or on the items making up the test, as in CRT. However IRT is controversial even within the quantitative framework, but is open to most of the same criticisms.

3. *Qualitative-Standards-Decontextualized*: The developmental mode of assessment, where the interest is in the growth of skills and concepts in themselves rather than in their applications, using pencil-and-paper testing rather than situated performances. Best used for finding the levels of understanding of basic concepts that students have attained so far; these levels could (and should) be incorporated into the curriculum as targets for instruction. Alternative framework research (White, 1988) and phenomenography (Marton, 1988) use situations and hierarchies specific to given topics, while SOLO uses a more general framework. Ordered-outcome testing, whether based on SOLO or any other growth model, also falls into this category. Backwash from such testing is likely to be helpful, as it encourages teacher and learner to "think higher".

4. *Quantitative-Measurement-Situated*: PA in NRT mode. This would involve setting tasks the grading of which is competitive; perhaps an athletics meet is an example. However, the assumption of the measurement model, that individual scores can be related to a stable trait, is really not what PA is about. For all intents and purposes, then, this is an empty cell.

5. *Quantitative-Standards-Situated*: many PA tasks could be scored and graded quantitatively, for example, number of correct applications or behaviors, use of rating scales. Such quantitative assessments are useful for communication or for manipulation in combining grades or determining cut-offs. While this may seem open to some of the objections already raised against quantitative assessment, it is much clearer to both teachers and learners that the numbers are used here for logistic purposes, and are not an implied statement about the quantitative nature of the performance. The form of understanding being created by the context is appropriate

to the real-world use of the knowledge in question precisely because it is situated.

6. *Qualitative-Standards-Situated*: "ecological" PA where the assessment task is situated and evaluated in a context close to the learning goals. Grading would be in terms of qualitative categories of mastery or competence. These might be quite task specific, or based on a more general framework such as SOLO. Best used for learnings that are meant to have direct real world applications, which accounts for much school learning but not all. Backwash is likely to be helpful for learning, as the testing situation should be designed to mimic the learning objectives, thus excluding cynicism and minimizing test-wiseness, short-cuts, or surface learning.

There are no doubt other modes of assessment that could be generated using different or additional dimensions. Perhaps one that is missing here is self- versus other-assessment; that is a particularly interesting pivot in the "situated" column. Nevertheless, the six, or rather five, cells generated here show a disproportion in terms of frequency of use in relation to the instructional context in which they are most useful.
Part of this disproportion is due a context that is inimical to the newer, alternative modes of assessment. The measurement establishment has in effect imposed a set of assumptions about the nature of educational measurement that is simply at odds with what should go on in classrooms; administrators have moreover encouraged this domination because the assumptions and practices of traditional measurement serve political and utilitarian administrative ends rather than educational ones (Wilson, 1994). Nowhere could this be clearer than in the current pressures towards competency-based testing currently being experienced in many countries (Biggs, 1994). The quantitative-decontextualized framework of assessment, whether CRT or NRT, has simply been hard to resist, either because of direct pressure, or more simply, from inertia.

Nevertheless, as the rapidly increasing interest and acceptance of PA shows, it is clear that the time has come for change, and increasingly teachers are recognizing this, perhaps more so in Canada (Bachor & Anderson, 1994) than in other countries. A real practical difficulty in the way of wider acceptance is not that practitioners see no need for change, but that they lack a framework for interpreting results and incorporating them into established grading schemes; the reliability and validity of the new techniques are not evident; while many forms of assessment and in particular report writing provide a heavy workload (Bachor & Anderson, 1994). The cost-benefits of PA are simply not seen to be favorable. All this suggests that much further work is necessary.
What needs to be done to make PA and qualitative assessment in general more acceptable?

Some Current Problems

Two main issues warrant discussion: determining conceptions of reliability and validity acceptable to practitioners; and adapting institutional structures to accept and make performance assessment workable in classroom contexts.

Conceptions of validity and reliability

As noted, the assumptions underlying PA and qualitative assessment in general are quite different from those underlying the measurement model. It is therefore inappropriate to apply the same tests and standards of reliability and validity. Messick's classic (1989)

review of validity set the tone for rethinking this issue, but this paper is not particularly oriented to qualitative assessment. However, his central point, that validity is not a property of the test, but of the interpretations and uses to which test scores are put, and their consequences, opened the way to considerable discussion and quite a good deal of consensus about reliability and validity specifically in relation to PA (Bachor, Anderson, Walsh, & Muir, 1994; Frederiksen & Collins, 1989; Haertel, 1991; Linn, Baker, & Dunbar, 1991; Messick, 1994; Moss, 1992, 1994; Shepard, 1993; Taylor, 1994; Wolf, Bixby, Glenn, & Gardner, 1991).

Perhaps the most fundamental insight is that validity is now seen as being grounded in the theory and knowledge base of learning and teaching, not in a science and technology of educational measurement, so that "in place of ranks, we will want to establish a developmentally ordered series of accomplishments" (Wolf, Bixby, Glenn, & Gardner, 1991: 63). The task of test construction then becomes rather different from the traditional model; one now requires a theory of learning, unfolding longitudinally, the construct being tested well represented in the elicited test behaviors and explaining them, and specific stepped targets, for each curriculum topic. The question is no longer: Does a test measure what it is supposed to measure? but: Does it do what it is supposed to do? (Shepard, 1993). Thus, assessment becomes contextualized, so that while accuracy, coverage or representativeness of teaching goals in the test items, and consistency are important, so too are fairness and adverse consequences, although whether these last questions belong in the domain of test validity, or professional ethics in general, is debatable, although their importance is not (Maguire, Hattie, & Haig, 1994).

Another fundamental shift from classic test theory is the role played by judgment and consensus, both in establishing construct validity (an end to those days when a test was "valid for anything with which it correlates" (Shepard, 1993)), and in interpreting test scores. Thus, while many are concerned about the low reliability of portfolios, for example (Bateson, 1994), it is important to recognize that the old additive model of compounding unreliability no longer applies. Specifically, a *hermeneutic* approach to drawing conclusions from test performances avoids this problem, as the aim is to arrive at a judgment by understanding the whole in light of the parts; it is not a case of judging single performances and then aggregating. An example is how a journal editor judges whether to accept or reject a manuscript on the basis of informed advice: even when rejecting a paper advocating a hermeneutic over an additive approach on the grounds that additivity is the accepted model! (Moss, 1994).

Finally, both fidelity or ecological validity and fairness require that multiple routes to the same goal performance are allowed. As in a music competition, performers choose their own particular items, often using different instruments, in order to show themselves at their best; analogically, then, students have the right to choose what to put in their portfolio (Moss, 1994). The final decisions, whether summative or formative, are holistic and qualitative, requiring expert judgment.

For educators brought up on traditional test theory these considerations are likely to feel alien, if not misleading. Undoubtedly Bateson (1994) is correct in saying: "University preservice training has generally failed dismally in preparing teachers for the testing, assessment and evaluation tasks they must undertake in their classrooms." (p. 239). However, there is now a growing theory and its technology that could inform

preservice and in-service teacher education that could replace the unsystematic but "amazing 'moccasin telegraph'" (*op. cit.*: 240) that spreads new and good ideas amongst teachers.

Institutional and systemic issues

The commitment and competence of teachers is certainly important in impeding change, but equally if not more important is the way institutions run. Educational institutions are *systems*, that is, a working whole made up of a set of component parts, each of which affects the other until the whole forms an equilibrium, that state of equilibrium becoming the system (von Bertalanffy, 1968). The prime example of a system is of course an ecological system, in which individual component organisms lie in a delicate state of symbiotic interdependence. However, systems theory can apply to almost any complex situation, including schools and classrooms (Yinger & Hendricks-Lee, 1993). In the classroom, the system comprises teacher, students, curriculum, teaching method, methods of assessment, and learning and teaching outcomes, all in a state of interdependence (Biggs, 1993). Part of the fierce resistance to change in assessment schemes is that assessment is so powerful a component, its backwash affecting the preceding chain of teaching and learning. Unless all participants are willing to change in the directions required, resistance will be considerable. The British Columbian situation (Bachor & Anderson, 1994; Bateson, 1994) illustrates both the difficulty and the possibility of change.

Reid (1987) points to three important components in the institutional system itself: the rhetoric, or the official aims of teaching; the technology, which would make possible the realization of these aims; and the social system of the institutions, which determines what is allowable within the institution. The social system comprises: the requirements established on a collegial basis, mostly informal but often given formal weight in Faculty meetings; the formal requirements of bureaucracy; and the requirements of the student body, which may be formal or informal. It is probably the social system, with its various collegial, accountability, and managerial agendas, that exerts most pressure on the assessment system in use (Biggs, in press).

The greatest pressures will be for using quantitative assessment. Most institutions require combining summative assessments at some stage: subunit results to obtain course results, course results to obtain year results. This puts almost irresistable pressure on teachers to use quantitative marking schemes, because marks are easily added up and averaged, and make discrimination between students extremely easy. Profiling or other qualitative schemes could be used, but usually are not in the event. Most teachers therefore mark quantitatively, with the sort of results we have discussed above (see Lohman, 1993).

Pressures towards decontextualised testing also exist in terms of convenience, security, and tradition. Institutionally, administrators feel it essential to have standardized procedures, timed testing conditions, and cheat-proof security, in the interests of fairness and accreditation, and in anticipation of possible law suits. The public, and students too, see a face validity in this. As for teachers, it is difficult enough breaking the habits of a lifetime to redesign situated assessment, let alone under the conditions prescribed by the bureaucracy.

Thus, changing the assessment system means setting up a new equilibrium, perhaps requiring a new technology, almost certainly requiring a new deal to be struck at all levels in the existing social system of the institution: with colleagues, with the bureacracy, and with students. It is not impossible, as the British Columbian experience shows, but it is likely to be difficult.

Conclusions

It is no exaggeration to say that the theory and practice of assessing learning are currently undergoing a major paradigmatic change. It is not a matter of CRT replacing NRT, but the intersection of a variety of movements, which interestingly have come from the broader educational canvas, not from the testing establishment itself. Certainly, the notion of CRT is coming to the forefront, but in connection with other views about: the qualitative nature of higher order learnings, and the situated nature of learning. Few of these ideas are particularly new, but their interaction becomes paradigmatic, suggesting heavily revized views of reliability and validity, and new formats of testing. The critical single notion underlying this is, amazingly enough, that educational considerations should drive testing, and not psychometric, bureaucratic, or political ones.

Assessment occupies a key place in determining quality learning outcomes, but assessment practices are part of a wider picture that includes but extends beyond the responsibility of any individual teacher. An institution is a holistic, interactive system, which for its own management has many procedures in place, with their own functional use. However, these procedures often determine teaching and assessment practices, which in turn influence students' perceptions of what and how they will learn. There are three main factors impeding change. The first and simplest is know-how; many teachers may simply not know how to improve their assessment techniques. Second, and more subtle, is the probability that they don't know that they don't know. If someone has a quantitative mind-set, really believing that we should teach, learn, and assess by numbers, then they won't even see that there is a problem. But finally, teachers have the institutional social system to deal with.

Obviously the road to better teaching, learning, and assessment is a complicated one that is beyond the control of educational researchers themselves. However, the history of assessment has shown the unfortunately negative effect the outdated measurement establishment has had on classroom practice. We at least can do something about that, and systems being what they are, perhaps we can then strike different and healthier equilibria than currently exist.

Note

This paper is based on an invited address to the Annual Meeting of the Canadian Association for Educational Psychology, Calgary, June, 1994. I am indebted to John Kirby, Bob Wilson, and one anonymous reviewer for their valuable feedback and advice.

References

Archbald, D.A. & Newman, F.M. (1988). Beyond standardized testing: Assessing

authentic achievement in the secondary school. Reston, Va.: National Association of Secondary Principals.

Bachor, D.G. & Anderson, J.O. (1994). Elementary teachers' assessment practices as observed in the Province of British Columbia. Assessment in Education, 1, 63-93.

Bachor, D.G., Anderson, J.O., Walsh, J. & Muir, W. (1994). Classroom assessment and the relationship to representativeness, accuracy, and consistency. The Alberta Journal of Educational Research, 40, 247-262.

Bateson, D.J. (1994). Psychometric and philosophic problems in "authentic" assessment: performance tasks and portfolios. The Alberta Journal of Educational Research, 40, 233-246.

Biggs, J.B. (1973). Study behaviour and performance in objective and essay formats. Australian Journal of Education, 17, 157-167.

Biggs, J.B. (1992a). Returning to school: Review and discussion. In A. Demetriou, M. Shayer, & A. Efklides (Eds.), Neo-Piagetian Theories of Cognitive Development (pp. 277-294). London: Routledge and Kegan Paul.

Biggs, J.B. (1992b). A qualitative approach to grading students. HERDSA News, 14(3), 3-6.

Biggs, J.B. (1993). From theory to practice: A cognitive systems approach. Higher Education Research and Development, 12, 73-86.

Biggs, J.B. (1994). Learning outcomes: Competence or expertise? Australian and New Zealand Journal of Vocational and Educational Research, 2, 1-18.

Biggs, J.B. (in press). Assessing learning quality: Reconciling institutional, staff and educational demands. Assessment and Evaluation in Higher Education.

Biggs, J.B. & Collis, K.F. (1982). Evaluating the quality of learning: The SOLO Taxonomy. New York: Academic Press.

Biggs, J.B. & Collis, K.F. (1989). Towards a model of schoolbased curriculum development and assessment: using the SOLO Taxonomy. Australian Journal of Education, 33, 149-161.

Biggs, J.B., Lam, R.Y.L., Balla, J.R. & Ki, W.W. (1988). Assessing learning over the long term: The `Ordered Outcomes' Model. Paper given to Annual Conference, Hong Kong Educational Research Association, 26-27 November.

Biggs, J.B. & Moore, P.J. (1993). The process of learning. Sydney: Prentice-Hall Australia.

Bloom, B.S., Hastings, J.T. & Madaus, G.F. (1971). Handbook of formative and summative education of student learning. New York: McGraw-Hill.

Brown, J.S., Collins, A. & Duguid, P. (1989). Situated cognition and the culture of learning. Educational Researcher, 18(1), 32-41.

Cizek, G. (1993). Rethinking psychometricians' beliefs about learning. Educational Researcher, 22(4), 4-9.

Cole, N.S. (1990). Conceptions of educational achievement. Educational Researcher, 19(3), 2-7.

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

Delandsheere, G. & Petrovsky, A.R. (1994). Capturing teachers' knowledge: Performance assessment. Educational Researcher, 23(5), 11-18.

Elton, L. & Laurillard, D. (1979). Trends in student learning. Studies in Higher Education, 4, 87-102.

Frederiksen, J.R. (1984). The real test bias: Influences of testing on teaching and learning. American Psychologist, 39, 193-202.

Frederiksen, J.R. & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Haertel, E.H. (1991). New forms of teacher assessment. Review of Research in education, 17, 3-29.

Hoffman, B. (1962). The tyranny of testing. New York: Collier.

Lai, P. & Biggs, J.B. (1994). Who benefits from mastery learning? Contemporary Educational Psychology, 19, 13-23.

Linn, R., Baker, C. & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

Lohman, D.F. (1993). Teaching and testing to develop fluid abilities. Educational Researcher, 22(7), 12-23.

Maguire, T.O. (1990). Grounded authentic assessment and teacher education. Paper presented to the Second Conference on Classroom Assessment, Vancouver, B.C., May 31 - June 1.

Maguire, T.O., Hattie, J. & Haig, B. (1994). Construct validity and achievement assessment. The Alberta Journal of Educational Research, 40, 109-126.

Marso, R.N. & Pigge, F.L. (1991). An analysis of teacher-made tests: Item-types, cognitive demands, and item construction errors. Contemporary Educational Psychology, 16, 279-286.

Marton, F. (1988). Describing and improving learning. In R.R. Schmeck (Ed.), Learning

strategies and learning styles. New York: Plenum.

Marton, F. Dall'alba, G. & Beaty, E. (in press). Conceptions of learning. International Journal of Educational Research.

Masters, G. (1987). New views of student learning: Implications for educational measurement. Research working paper 87.11. University of Melbourne: Centre for the Study of Higher Education.

Masters, G.N. & Hill, P.W. (1988). Reforming the assessment of student achievement in the senior secondary school. Australian Journal of Education, 32, 274-286.

Messick, S.J. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Messick, S.J. (1994). The interlay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. Review of Educational Research, 62, 229-258.

Moss, P.A. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

Newman, F.M. & Archbald, D.A. (1992). The nature of authentic academic achievement. In A.R. Tom (Ed.), Toward a new science of educational testing and achievement. Albany: State University of New York Press.

Popham, W.J. (1987). The merits of measurement-driven instruction. Phi Delta Kappan, 68, 679-682.

Popham, W.J. & Husek, T.R. (1969). Implications of criterion-referenced measurement. Journal of Educational Measurement, 6, 1-9.

Ramsden, P. (Ed.) (1988). Improving Learning: New Perspectives. London: Kogan Page.

Reid, W.A. (1987). Institutions and practices: Professional education reports and the language of reform. Educational Researcher, 16(8), 10-15.

Scarino, A., Clark, J. & Brownell, J. (1994). A framework for target oriented curriculum renewal in Hong Kong.

Shepard, L.A. (1991). Psychometricians' beliefs about learning. Educational Researcher, 20(6), 2-16.

Shepard, L.A. (1993). Evaluating test validity. Review of Research in Education, 19, 405-450.

Tang, K.C.C. (1991). Effects of different assessment methods on tertiary students'

approaches to studying. University of Hong Kong: Unpublished Ph.D. Dissertation.

Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large scale assessment reform. American Educational Research Journal, 31, 231-262.

Trigwell, K., Prosser, M. & Taylor, P. (1994). Qualitative differences in approaches to teaching first year university science. Higher Education, 27, 75-84.

Von Bertalanffy, H. (1968). General systems theory. New York: Braziller.
White, R.T. (1988). Learning Science. Oxford: Basil Blackwell.

Wiggins, G. (1989). Teaching to the (authentic) test. Educational Leadership, 46, 41-47.

Wilson, R.J. (1994). Back to basics: A revisionist model of classroom-based assessment. Presidential Address, Annual Meeting, Canadian Educational Researchers Association, CSSE, Calgary, Alberta, June.

Wilson, R.J. & Kirby, J.R. (1944). Introduction: Special issue on cognition and assessment. The Alberta Journal of Educational Research, 40, 105-108.

Wolf, 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.

Wong, C.S. (1994). Using a cognitive approach to assess achievement in secondary school mathematics. University of Hong Kong: Unpublished M.Ed. Dissertation.

Wright, B.D. & Stone, M.H. (1979). Best test design: Rasch measurement. Chicago: MESA press.

Yinger, R.J. & Hendricks-Lee, M.S. (1993). An ecological conception of teaching. Learning and Individual Differences, 5, 269-282.

TABLE 1: DIMENSIONS AND MODES OF ASSESSMENT

		Context (examples)	
	Model		
Framework		Decontextualized	Situated
Quantitative	Measurement	1. NRT (MC test)	4. PA (NRT) (… ?)
	Standards	2. CRT (Mastery Learning)	5. PA (CRT)
Qualitative	Standards	3. SOLO (Developmental)	6. PA (Ecological)