The Three Roles of Assessment: Gatekeeping, Accountability, and Instructional Diagnosis

Philip Nagy

External assessment began as an imposed activity to provide quality control for a process. This gatekeeping role has a norm-referenced focus. A second role, ensuring accountability, emerged to judge the quality of education, an activity that has remained essentially norm-referenced. A third role, instructional diagnosis, is a recent phenomenon driven by the need to improve education and to provide educational, not political, justification. Traced from a measurement perspective, two requirements must be met for external assessment to yield instructionally diagnostic information: a reconceptualization of reliability, and development of more detailed and facilitative mechanisms for test scoring and interpretation.

L'évaluation externe peut jouer plusieurs rôles : un rôle de garde barrière comportant une approche normative; un rôle relié au principe de redevabilité jugeant de la qualité de l'éducation en termes politiques; un rôle de diagnostic pédagogique répondant au besoin d'améliorer l'éducation du point de vue de l'apprenant. Dans chacun de ces cas, deux exigences doivent être remplies pour que l'évaluation externe fournisse de l'information diagnostique sur le plan pédagogique : une reconceptualisation de la notion de fiabilité et l'élaboration de mécanismes plus détaillés et utiles pour la correction des tests et l'interprétation des résultats.

Assessment has three roles or functions. The role with the longest history is that of gatekeeping (National Commission on Testing and Public Policy, 1990), in which assessment determines who is granted a privilege such as admission or graduation. To this role has recently been added the role of ensuring accountability (Darling-Hammond & Ascher, 1991; Wohlstetter, 1991), in which assessment is used to decide if schools are working well. The third role is that of instructional diagnosis (Levesque, Bradby, & Rossi, 1996), in which assessment is used to find out what students do and do not know, and what to do about it. These three roles have not simply replaced each other; rather, additional requirements have gradually been added to the expectations held for external assessment. In any given assessment program, these roles may overlap. The purposes to be met in each role, however, need to be clearly distinguished, as program design and procedures need to be compatible with their purposes.

CANADIAN JOURNAL OF EDUCATION 25, 4 (2000): 262–279

In this article, I argue that the roles of gatekeeping and ensuring accountability call for norm-referenced assessment, resulting in comparisons at the expense of guidance for instruction. Even where standards of proficiency are set, such as with programs in Alberta and Ontario (Alberta Learning, n.d.; Education Quality and Accountability Office, 2000a), the major interpretive tool is comparison – whether X had more students meet the standard than did Y, or whether X had more students meet the standard this year than last. Schools do indeed improve test scores given only such comparisons. This is done by increasing instructional emphasis on the constructs used for reporting, such as *numeration* or *problem solving*. The criteria for judging success, nevertheless, remain norm-referenced, focussing on whether a school's *relative* standing improves. A difficulty, however, is that, "As a result of the accountability emphasis, policy relevant variables are neither collected nor integrated with test results, so that educational practices and policies are not easily linked to outcomes" (Cooley, 1991, p. 3). The hope is that external pressure will cause improvement, although there is evidence that teachers do not always know what to do about low achievement (Madaus & O'Dwyer, 1999). Diagnosis based on accountability data is difficult; such data provide the motivation, but not the means, for improvement.

Teachers already engage in diagnosis, but they do so on the basis of informal observation and results from teacher-made tests administered for this specific purpose, rather than from external tests. Even formal teacher-made tests are often summative, given to gather evidence for reporting rather than to plan for instruction. Coffman (1993) has argued that only teachers and not external testing can provide diagnosis. Serafini (2001) agrees, arguing that external tests come from an acquisition model of learning, and teacher diagnosis from a constructivist model. I argue that although gatekeeping and ensuring accountability are ascendant, improving instruction is of greater benefit. And though I recognize that a single test cannot accomplish everything, I argue too that assessment *programs* need to evolve to include the benefits of instructional diagnosis, which will provide greater justification for their substantial costs.

ASSESSMENT AS GATEKEEPING

History

Public education began about 1850 in eastern Canada, and slightly later in the west. Earlier, education was a privilege for the elite, left to tutors and private schools. As society developed, the need for educated people grew. Public primary education became first accessible, then, eventually,

compulsory. Quality was uneven; academic requirements for teachers were slow to develop, and availability of teacher education did not meet demand. Although high school was still for the elite, by the 1930s and 1940s it increasingly came to be seen as necessary and gradually became more common. However, because children of the poor had to contribute to the family income, they often went to school only when convenient (Brown, 1999).

In this context, there arose a system of quality control based on government exams, a centuries-old tradition. These exams were intended to ensure a common standard and provide a sense of fairness. They were and are expensive, and their popularity was and remains tied to availability of resources. As *product* control, they are most popular when there are difficulties with *process* control, that is, a lack of sufficient numbers of good teachers and well-equipped schools. During the first decades of public education in Canada, examinations were needed because education was still something of a privilege, and society begrudged the cost of schools for the poor (Brown, 1999). Starting in the mid-1950s, society began to tolerate and even require greater spending on education (Gidney, 1999). As standards rose and became more uniform, external exams became less popular. For example, Ontario dropped high-school entrance exams in the 1930s. Many, but not all, provinces also dropped external, high-school, graduation exams. Public money for public education was readily available, and ensuing differences in standards across schools were tolerated because there was an ability and willingness to pay for all.

Fairness has two meanings, and a common standard represents only the narrower one. If two students with varying backgrounds and differently qualified teachers write the same exam, the resulting fairness is that of a common standard rather than the broader fairness of equal opportunity. At one time, the province of Newfoundland suffered from substantial urban-rural disparities in the quality of education. Rural teachers took one of two different positions in response to this disparity: some evaluated hard, to instill a sense of standards, and others evaluated generously, to maximize the students' chances of making something of their lives. Both positions are arguably sensible (Nagy, 1984), but they appeal to different meanings of fairness.

Recent cutbacks in educational spending, indeed in all government spending, bring this dichotomy to the forefront. When education became public, it shifted from a means of preserving the social structure to a means of changing it. Paid for by those with income or property, it was a way out of poverty for the less fortunate. The current trend towards spending cutbacks and a "new right" accountability contribute to a reversal of this

ideal, if not in intention, then certainly in effect. The lack of public willingness to pay for education has returned. Gidney (1999) links this to the current political climate, noting "the role of the school in a capitalist system devoted to the reproduction of social economic inequality" (p. 182). When a refusal to provide resources for schools is set beside a renewed demand for external examinations, we run the danger of replacing the broader fairness of equal opportunity with the narrower fairness of a consistent standard.

Measurement Theory and Gatekeeping

How has measurement evolved with and contributed to the gatekeeping role? The idea of selecting some individuals from others has been the impetus for the development of norm-referenced test theory. For example, we accept now that test scores have less than perfect reliability. We acknowledge, and try to minimize, a small number of false positives and false negatives in the certification of students. Beyond this, reliability, as generalizability theory (e.g., Brennan, 1997), has evolved from providing estimates of total error in a test score to providing estimates of the relative importance of different sources of error. We can tell, for example, whether we will get a more accurate estimate of writing ability by having a student write two passages, each scored by three judges, or three passages, each scored by two judges.

Recent research on performance assessment, including written exams, has revealed the problem of student-task interaction (Ruiz-Primo, Baxter, & Shavelson, 1993). Two different but content-equivalent exams may produce different grades, because a student may emphasize, in exam preparation, some poems, battles, or experiments rather than others. Any given question is but one of several possible equivalents. The generalization from *this task* to *tasks of this type* is dubious. Although average difficulties of exams set in different years can be equated, resulting exams still put students in different rank orders. This is rarely acknowledged; test content is taken to be *the* domain of interest rather than merely a sample of areas of interest.

Methods for maintaining uniform exam difficulty over time have improved with the development of item response theory (Hambleton, Swaminathan, & Rogers, 1991). Apart from setting longer exams (e.g., asking about all poems), however, little can be done about student-task interaction if all students write the same exam. Although computer-adaptive testing is an available solution, the cost of implementing this in schools seems prohibitive in a time of decreasing budgets.

ASSESSMENT AS ACCOUNTABILITY

History

The second role of assessment is to ensure accountability, that is, to judge and improve the quality of education. Canada has a shorter history with assessment for accountability than does the U.S.A., where greater emphasis on local control has given rise to earlier and stronger efforts by central authorities to impose quality control. Linn (2000) notes four historical stages of U.S. accountability: *program*, whether instructional intentions are realized (1960s); *minimum competency*, whether all students are reaching a minimum standard (1970s); *school and district*, how jurisdictions compare (1980s); and *standards-based*, whether students are meeting specific goals (1990s). Although the Canadian experience is more compressed, we have some familiarity with each of these stages.

Until about 1965, it was common to see almost-adult youth sitting in Grade 8 waiting to be old enough to leave school. If a student could not do what the school asked, this was viewed as the student's fault. Prior to the mid-1960s, education centred on academics. The mid-1960s saw dramatic changes with the introduction of vocational training and expansion of the college system. Many non-academic courses were introduced, and the curriculum and associated services became more complex and costly. Demand increased and enrolment soared. Schools began to serve much more diverse populations and were called on to carry out functions, some would argue, that traditionally belonged to family, community, or church. Huge increases in population, from the baby boom, and, in cities, by large-scale immigration (Gidney, 1999), exacerbated these changes. This last type of growth also brought with it increased demands for language training. Salaries for teachers rose, and school boards hired increasing numbers of support staff.

In this time of ample resources, most provinces that had provincial exams at the end of high school dropped them; this was particularly so during the late 1960s and early 1970s. Subsequently, concerns about grade inflation and about quality and increasing costs led to these exams being re-introduced in Alberta and British Columbia. Although investigation showed grade inflation in Ontario to be a simple scale shift (Traub, Wolfe, Wolfe, Evans, & Russell, 1977), some people argued for the return of public exams (Coalition for Education Reform, 1994). They claimed to remember a better time, when students were superior and everyone who graduated had sufficient skills to enter university or college, or to be gainfully employed. Ontario has introduced a literacy test as a high-school graduation requirement.

This period of changing school mandates, increasing student numbers, changing student types, and consequent increasing costs raised concerns about quality and expenditures, and ushered in an era of accountability. Provinces and school systems felt pressure to compare their achievement to that of others. This trend continues to the present.

The stakes involved in accountability range from low to high. Low stakes include publication of school results and requirements to develop public improvement plans (Alberta Learning, n.d.; Education Quality and Accountability Office, 2000b) and inclusion of results in school accreditation procedures (British Columbia Ministry of Education, n.d.). Higher stakes include enhancement of school budgets and even payments to individual teachers (California Department of Education, n.d.; Pennsylvania Department of Education, n.d.). Although these examples suggest that high-stakes testing is a U.S. phenomenon, Ontario has announced that school funding will be linked to achievement data (Ontario Ministry of Education and Training, 2001). Unlike some U.S. jurisdictions that have punished low-achieving schools financially, Ontario has chosen to offer additional funding to such schools.

The work of Ontario's Education Quality and Accountability Office (EQAO) exemplifies the issues faced in demonstrating accountability to the public. Like similar agencies, EQAO bases its accountability evidence on results of curriculum-embedded performance assessment material, with a multiple-choice component. The EQAO uses a 4-point scale to report overall performance in reading, writing, and mathematics, and in a number of categories or strands within each of these.

A major task of such agencies is to educate the media and schools about the meaning of assessment data. Because of the efforts of these agencies, the quality of public discourse on this issue is improving. Comparison by rank order is becoming less acceptable. The role of error and confidence intervals has been recognized, and the link between socioeconomic conditions and achievement (e.g., Nagy, Traub, & Moore, 1999; Payne & Biddle, 1999) has become more visible.

In the Third International Mathematics and Science Study (TIMSS) (Robitaille, Taylor, & Orpwood, 1996) Canada ranked in the middle of the developed nations. Critics looked at our results and those of economically more successful countries and concluded that we could blame the education system for our economic woes.

There are problems with this conclusion. This judgement downplays the fact that Canadian industry has a poor record of training its workforce (Nagy, 1996). It is also noteworthy that when the Japanese economy slowed in the 1990s, public discussion of possible reasons did not centre on a sudden deterioration of Japanese education.

The atmosphere of concern about schools has provided a forum for some self-serving rhetoric from those who simply wish to cut taxes and privatize schools, the spectre of unwillingness to share reincarnated. For example, Ontario's results were almost exactly in the middle of those from the more than 40 jurisdictions involved in TIMSS. However, the Ontario Ministry of Education and Training (1997) circulated a graph showing results for only the top half of these countries, misrepresenting Ontario's performance and allowing the interpretation that Ontario was at "rock bottom."

International tests, despite their susceptibility to such misuse and misinterpretation, have done some good. They showed that mathematics curricula in Canada were "a mile wide and an inch deep," as compared with, for example, the Japanese curriculum. Teachers did too much skill-and-drill, and not enough teaching for understanding. Like children in many other countries, those in Canada receive little exposure in elementary school to teachers who know much about mathematics or even like it.

Measurement Theory and Accountability

Measurement issues with respect to accountability are more complex than those in the gatekeeping context. One central question is the trade-off between validity and reliability, or between curricular relevance and accuracy. Those who conduct assessments use a combination of multiple-choice items, written extended response formats, and, occasionally, non-written performance assessments (e.g., laboratory exercises). Extended responses are more able to tap a broader range of skills and objectives, and they give better curricular signals (Archbald & Newmann, 1988) than do multiple-choice items. On the other hand, multiple-choice items deal better with the generalization from *this task* to *tasks of this type*. Further, multiple-choice items are more reliably scored, at a lower cost than written or performance items. They also make it easier to equate tests over time (Hambleton, Swaminathan, & Rogers, 1991).

The major issue concerning reliability of performance assessments is that discrimination among students is often not the primary goal, so that traditional reliability is lower. For example, extrapolating from Wolfe, Wiley, and Traub (1999), the standard error of scores in the Ontario assessment program is about 0.7 on the 4-point scale used, meaning that individual scores contain substantial error. These assessments, not intended to discriminate, show low estimated accuracy using traditional methods, and this in turn calls into question the appropriateness of such estimates.

The perceived lower reliability of extended-response assessments has raised the question of whether our conception of reliability is appropriate for such tasks. Some researchers have called for a reconceptualization of

reliability (e.g., Worthen, 1993), and Messick (1995) has referred to "a trade-off between the valid description of the specifics of a complex task performance and the power of construct interpretation" (p. 7). Hambleton (2000) recently edited a special issue of *Applied Psychological Measurement* that deals with this problem from the perspective of traditional measurement theory.

Another major issue in accountability assessment is whether to administer the same assessment instrument to every student or different samples of items to different samples of students. Test developers can solve the problem of generalization from a small number of tasks by using matrix sampling, a technique in which different samples of students write different but parallel tests (this is not possible in traditional gatekeeping situations). This procedure also provides a mechanism for piloting new items, thus allowing the assessment to evolve with the curriculum. The disadvantage of matrix sampling is that, although it produces good information for larger schools and for school districts, it makes the estimation of scores for individuals and smaller schools more difficult and less accurate.

Although administering the same test to all students provides school-level results, it raises concerns about test security, and the problem of generalization from *this task* to *tasks of this type* remains. Thus the question arises: how important are school-level test results? The answer can be seen in results of the first Ontario assessment. The EQAO program tested all students in Grade 3 but only a sample in Grade 6. The Grade 3 results, available for every school, were clearly noticed. In contrast, the Grade 6 results in mathematics went virtually unnoticed by most educators. Without the impetus of every-school results, it seems, external assessments have little effect.

Proponents of professional models of accountability (e.g., Darling-Hammond & Ascher, 1991) argue that teachers, properly supported, will act on good assessment data without coercion. On the other hand, Madaus and Kellaghan (1993) argue that there must be moderately high stakes to achieve any kind of impact. Black (1994) has outlined the problems of doing assessment in the face of difficult relationships between teachers and government. There are many parallels between the British experience he reports and current conditions in parts of Canada. Certainly, cooperation of teachers is fundamental to non-coercive use of assessment data.

Much work remains with respect to accuracy of scores, including our conception of it. However, if current conceptions of reliability hold sway, and if the public insists on comparisons, such comparisons need to be accurate. Rolling averages over a few years or some similar mechanism will provide a better indicator of school achievement. Achievement needs to be set in a value-added (Meyer, 1997) or socioeconomic (Nagy, Traub, &

Moore, 1999) context. The goal of assessment, however, should be more than comparisons.

ASSESSMENT AS INSTRUCTIONAL DIAGNOSIS

History

Because large-scale assessment for instructional diagnosis is more promise than reality, there is little history to report. Mehrens (1998) notes that most writing on effects of assessments is data-free rhetoric. He further points out the difficulties of doing research that would prove the effects of assessment on instruction. Based on limited evidence, he concludes that if the stakes are high enough and teachers consider the material assessed to be appropriate, they will shift instruction to cover test content. If not, the impact of the assessment will not be obvious. Additionally, there are reports of influences towards inappropriate teaching to the test (Cannell, 1988; Madaus & Kellaghan, 1993) and the effects of teaching of test-taking skills (Bangert-Drowns, Kulik, & Kulik, 1983; Mehrens & Kaminski, 1989; Popham, 1991).

There are reasons why it is difficult to apply large-scale assessment results to instructional diagnosis. Serafini (2001) argues from a constructivist viewpoint that top-down accountability is so fundamentally different from providing data for instructional diagnosis that educators should not expect any classroom effects. He goes on to advocate replacing large-scale assessment with "assessment as inquiry" (p. 387), much in line with Darling-Hammond and Ascher's views (Darling-Hammond & Ascher, 1991). Such assessment would be done in the classroom and would focus on determining specific reasons for student misunderstanding or lack of skill. He argues as if teachers had some say in whether they engage in top-down accountability practices; in a climate of top-down accountability, this is not the case.

There is compelling evidence on teachers' inability to understand and use assessment data. Corcoran and Goertz (1995) present an example of high-stakes assessment to improve science scores when teachers simply do not know the science. They also note that professional development is often focussed on short-term behavioural changes rather than long-term learning needs. Firestone, Mayrowetz, and Fairman (1998) studied the influence of performance-based assessment in mathematics in Maryland and Maine. In agreement with Mehrens (1998), they found some effort to align the curriculum with the tests but little evidence that basic instructional strategies changed. They concluded that teachers need to learn a considerable amount if they are to change their instructional practices.

In a report of imposed performance assessments in England, Wales, and Scotland, Madaus and Kellaghan (1993) noted the need for high stakes if any real changed was to be expected. At the same time, they reported the deleterious effect of comparisons. English and Welsh results were reported in the popular press in the form of "league tables" comparing schools; this was not done in Scotland. The stress caused by the comparisons overwhelmed any potential instructional uses: teachers in England and Wales reported that the assessments were more disruptive than did teachers in Scotland.

Measurement and Instructional Diagnosis

Accuracy of Performance Assessments

When evidence appeared about the low reliability of performance assessments (e.g., Ruiz-Primo, Baxter, & Shavelson, 1993), a debate ensued in the measurement literature about the conception of reliability. Worthen (1993) characterized the positions in this debate as those who would insist on the same high standards as measured in traditional ways versus those who would abandon or rethink traditional reliability criteria.

For the argument favouring rethinking, the 1970s debate over reliability of criterion-referenced testing is important as a precedent. It led to a conceptual breakthrough in how reliability was understood and, eventually, to the incorporation of the idea of decision consistency in generalizability theory (Brennan, 1997). A similar breakthrough in the measurement community may be required to gain widespread acceptance of the accuracy of performance assessment in large-scale assessment contexts.

Delandshere and Petrosky (1994, 1998) argue that in complex domains, the idea that a performance task is a sample from a domain does not apply, and thus traditional reliability does not operate. They suggest replacing the traditional conception of reliability with *confirmation*, noting that "Confirmation as we have proposed it seems also to blur the distinction between the measurement notions of reliability and validity" (1994, p. 17). Moss and her colleagues (Moss, 1994; Moss et al., 1992) have taken up the idea of reliability as a subset of validity, arguing for the possibility of validity without traditional reliability. They offer "alternative procedures and criteria for investigating validity in these less standardized domains" (Moss et al., 1992, p. 12).

Haertel (1999) notes that validity has long been regarded as a process of argument rather than a calculation, and Kane, Crooks, and Cohen (1999) suggest that reliability, as well as validity, might also require a process of argumentation in place of or in addition to a simple calculation.

The discussion of the meaning of reliability is largely conceptual, but two developments can be construed as technical. Moss (1994) puts forward hermeneutic approaches to an alternative conceptualization of reliability. Traditional reliability is decontextualized, and hermeneutics is holistic. She offers as examples the conferring of graduate degrees and the academic hiring process, where participants take disagreement seriously and try to resolve it, rather than treating it as error. Her proposal is more in line with Darling-Hammond and Ascher's professional model of accountability (Darling-Hammond & Ascher, 1991) than with an externally imposed system.

Nichols and Smith (1998) offer a technical contribution to the reconceptualization of reliability, in the form of an extension of generalizability theory. They point out that traditional views of reliability are embedded in a trait view of learning. They provide an example involving trait and constructivist views of writing ability, and argue using data from the National Assessment of Educational Progress that the trait model of writing is untenable. "In contrast, reliability analyses that assume that multiple strategies or schemas are used by different test takers to respond to different assessment tasks across different occasions might conclude that cognitively complex assessments adequately meet reliability standards" (p. 30). In a reliability system that uses complex cognitive models, variance associated with different groups of students with different experiences might be either included as true score variance or excluded entirely from the analysis. "As a result, reliability studies of alternative assessments probably provide inflated estimates of the amount of error involved in the measurement procedure" (p. 32).

It appears that mining performance data for diagnostic value requires a more widely shared acceptance that, for this purpose, the data are accurate enough when aggregated to the group (classroom, school, or district) level.

The Difficulty of Instructional Diagnosis

There is a line of thought that testing itself causes improvement in achievement. This may be true in a few schools, and it may be true of the first, easily accomplished gains in achievement. However, Covington (1996) has dubbed this the "myth of intensification" (p. 24), asserting that it is much easier to raise test scores the first few percentage points than the second or third; and that it is easier for some (e.g., the most motivated) schools to improve than for all schools to improve. Comparative results alone, though they offer motivation, and ammunition for debate, do not actually provide much help per se. Diagnosis that is based upon instructional consequences

of the implications of assessment results is needed for sustained improvement across large numbers of schools.

How difficult is this? In the Ontario assessments, EQAO gives to schools summaries of the proportion of students at each of four levels of performance, and districts receive school-level summaries. Schools must then design improvement plans. Examples of actual plans produced in response to this requirement (EQAO, 2000a, 2000b) are good, sensible, and quite detailed, but they are generic educational plans. With nothing but general indicators of success that reflect a fixed point in time within one grade, on such global constructs as *Number Sense and Numeration, Problem Solving,* and *Reasoning,* how could suggested strategies for student improvement be anything but generic?

Faced with low performance in the 1999 mathematics categories *Problem Solving* and *Communication,* one district suggested that, "To improve, students need to use appropriate and innovative strategies to solve problems, use correct mathematical terms, and give clear and precise explanations to justify reasonableness of solutions" (EQAO, 2000a, p. 8). Perhaps the second of these suggestions is ready to implement, but the other two are no more specific than a generic curriculum document. Similarly, suggested strategies to improve reading scores often include implementing home-reading programs, purchasing materials, and sending teachers to professional development sessions. Did schools and districts need data to generate such strategies? This is neither a criticism of EQAO nor of school districts; rather, it is a comment on the difficulty of extracting diagnostic information from large-scale assessment scores. There is little in the assessment results specific enough to provide any kind of diagnosis.

Possibilities for Progress in Instructional Diagnosis

How do we progress? There may be promise in development of differential and more detailed scoring systems. In an assessment system involving huge numbers of students, and with the need for quick reporting, scoring has to happen quickly, with low-inference, non-speculative procedures. However, detailed analytic scoring and higher-inference scoring procedures may yield better information about student thinking.

The scoring of performance assessments by classroom teachers or central panels is based on rubrics, which arise from assessment-based instruction (Andrade, 2000). A body of literature on rubrics-based teaching focuses on using teacher-constructed rubrics rather than centrally constructed scoring. Teacher-constructed rubrics are better referenced to instruction since they are "closer" to what is happening in the classroom. This observation

follows Serafini's (2001) notion of assessment as enquiry, and Moss et al.'s (1992) suggestions for accountability as an auditing process grounded in classroom assessment. A key issue related to such adaptation, however, includes the challenge of translating a necessarily generic scoring rubric from a large-scale assessment into an instructionally specific rubric for classroom use (Arter, 1999; Popham; 1999); this requires a thorough understanding of rubric design, which, generally speaking, teachers do not yet have.

Two areas need further discussion: one is technical, concerning improved links between rubrics-based assessment and teaching practices; the other is political, concerning acceptance of accountability systems that give instruction a higher priority than reporting to the public. Elements of such discussions exist. Arter (1999) has recently written on teacher-developed rubrics, arguing that instructional rubrics blur the distinction between instruction and assessment and emphasize evidence of good thinking. This latter point is consonant with much current discussion in the measurement literature of validity of performance assessment (Kane, Crooks, & Cohen, 1999; Popham, 1999). Arter also speaks of omitting from rubrics descriptors of performance that do not distinguish good from poor quality, and of focussing on generalized rather than task-specific criteria, two ideas in harmony with the discussion of a reconceptualized reliability I outlined earlier. Rubrics require development as both instructional tools and devices capable of reasonably accurate assessment. There seems no a priori reason why these two cannot be worked on at the same time.

Much effort is needed to move the focus of accountability from outside the classroom to inside. Hart (1994) says that assessment with instructional utility has "systemic validity" (p. 13), and although additional terminology to confuse discussions of validity is unnecessary, her admonition to "avoid the temptation to import and implement a ready-made assessment program without extensive teacher consultation" (p. 86) is well taken. This goes to the heart of the top-down versus bottom-up approaches to accountability (Darling-Hammond & Ascher, 1991).

Harris and Carr (1996) have written about the problem of development of a standards-based curriculum, and the importance of carrying local standards from the school to the wider community. This approach is the opposite of imposing external standards on the school, and considerable time and effort, as well as a major climate change, will be required for it to gain widespread acceptance in the measurement community.

The gap between those who wish to apply traditional reliability criteria to performance assessment, and those who wish to supplant them, may be narrowing. Three articles in a special issue of *Applied Psychological Measurement* on performance assessment (Hambleton, 2000) make relevant points.

Clauser (2000), building on the work of Kane, Crooks, and Cohen (1999), raises key concerns about scoring of performance assessments: what aspects to score, what criteria to use, and how to develop and apply these criteria. Brennan (2000) raises the possibility of using different scoring rubrics as facets of a generalizability analysis, although he does not refer to Nichols and Smith's (1998) work. Finally, Miller and Linn (2000) offer the intriguing idea that high stakes could lead to instructional alignment or other classroom practices that reduce the person-task interaction. This notion is certainly worth empirical investigation.

FURTHER OBSERVATIONS

I have already discussed longer-term solutions to the problem of gleaning instructional diagnosis from assessment data. In the shorter term, what might be done? As a start, assessment programs should follow Cooley's (1991) advice and collect "policy relevant" (p. 3) variables to help interpret achievement data.

One positive aspect of the experiences of Ontario and British Columbia has been the professional development of teachers involved in item development and scoring. This practice suggests a mechanism for the promotion of performance-based assessment. If school districts can provide opportunities for large numbers of teachers to develop these skills, it will increase the chances of using accountability to smooth the path of learning. Educators may find that teachers' involvement in the development and scoring of assessment instruments at the classroom level may become as important for instructional improvement as the actual test results.

Accountability as presently conceived seems forced on educators, and public thinking is norm-referenced, despite the criterion-referenced nature of performance assessments. Comparison alone is not particularly helpful; in fact, when relations between government and teachers are poor, it is distinctly unhelpful. On the other hand, no one ought to, or actually does, rely on external assessment for individual diagnosis. As Coffman (1993) has noted, individual diagnosis is a within-classroom task for the teacher. In practice, many average and better students get along well enough without a lot of individual diagnosis. Were this not so, the classroom job would be unmanageable. The question is simply how to make accountability results more helpful, so that their costs can be better justified.

In many ways, measurement theory seems caught halfway in the evolution from a framework focussed on norm-referencing and multiple-choice items to one based on a performance-assessment framework that does not emphasize discrimination. Educators have abandoned the accuracy of more

traditional tests, but that accuracy, as measured by traditional means, is under question. The early 1970s saw the development of generalizability theory, which in turn allowed a reconceptualization of traditional reliability to fit criterion-referenced testing. What educators need at this time is a reconceptualization of traditional reliability to fit performance assessment. This change has started, and within a few years, the conception of accuracy may have evolved so that a number of the more pessimistic claims about accuracy are no longer taken as true.

Assessment means quality control. In addition to the existence of traditional public exams, there is a worldwide trend towards comparative quality control, driven in large measure by the globalization of our economies. Accountability is costly, and as currently conceived, does not appear to hold much benefit for teachers and students. In fact, the comparative aspects seem to hinder positive instructional uses of the data. The task ahead is to develop ways of using methods for product control to improve the process.

ACKNOWLEDGEMENTS

I thank Todd Rogers, University of Alberta, for comments on earlier drafts of this article.

REFERENCES

Alberta Learning (n.d.). *Provincial testing.* Retrieved October 2, 2002 from http://www.learning.gov.ab.ca/k_12/testing/

Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), 13–18.

Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school.* Reston, VA: National Association of Secondary School Principals.

Arter, J. (1999). Performance criteria: Integrating assessment and instruction. *The High School Magazine, 6*(5), 24–28.

Bangert-Drowns, R., Kulik, J., & Kulik, C. (1983). Effects of coaching programs on achievement test performance. *Review of Educational Research, 53,* 571–585.

Black, P. J. (1994). Performance assessment and accountability: The experience in England and Wales. *Educational Evaluation and Policy Analysis, 16,* 191–203.

Brennan, R. L. (1997). A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice, 16*(4), 14–20.

Brennan, R. L. (2000). Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement, 24,* 339–354.

British Columbia Ministry of Education. (n.d.). *Accreditation program for schools.* Retrieved May 29, 2002 from http://www.bced.gov.bc.ca/accreditation/

Brown, R. (1999). *A study of absenteeism in the Toronto Board of Education.* Unpublished doctoral dissertation, University of Toronto.

California Department of Education. (n.d.). *Public Schools Accountability Act of 1999: Awards program.* Retrieved October 2, 2002 from http://www.cde.ca.gov/psaa/awards

Cannell, J. J. (1988). Nationally normed elementary achievement testing in America's public schools: How all 50 states are above the national average. *Educational Measurement Issues and Practice, 7*(2), 5–9.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24,* 310–324.

Coalition for Education Reform. (1994). *Could do better: What's wrong with public education in Ontario and how to fix it.* Toronto: Coalition for Education Reform.

Coffman, W. E. (1993). A king over Egypt, which knew not Joseph. *Educational Measurement: Issues and Practice, 12*(2), 5–8, 23.

Cooley, W. W. (1991). State-wide student assessment. *Educational Measurement: Issues and Practice, 10*(4), 3–6, 15.

Corcoran, T., & Goertz, M. (1995). Instructional capacity and high performance schools. *Educational Researcher, 24*(9), 27–31.

Covington, M. V. (1996). The myth of intensification. *Educational Researcher, 25*(8), 24–27.

Darling-Hammond, L., & Ascher, C. (1991). *Creating accountability in big city schools* (Urban Diversity Series No. 102). New York: National Center for Restructuring Education, Schools, and Teaching. (ERIC Document Reproduction Service No. ED334339)

Delandshere, G., & Petrosky, A. R. (1994). Capturing teachers = knowledge: Performance assessment. *Educational Researcher, 23*(5), 11–18.

Delandshere, G., & Petrosky, A. R. (1998). Assessment of complex performances: Limitations of key measurement assumptions. *Educational Researcher, 27*(2), 14–24.

Education Quality and Accountability Office. (2000a). *Educator handbook* (2nd ed.). Toronto. Queen's Printer for Ontario.

Education Quality and Accountability Office. (2000b). *Ontario Report and Guide on School Improvement Planning 1999–2000.* Toronto: Queen's Printer of Ontario. Retrieved October 2, 2002 from http://www.eqao.com/eqao/home_page/pdf_e/00/00P056e.pdf

Firestone, W. A., Mayrowetz, D., & Fairman, J. (1998). Performance-based assessment and instructional change: The effects of testing in Maine and Maryland. *Educational Evaluation and Policy Analysis, 20,* 95–113.

Gidney, R. D. (1999). *From Hope to Harris.* Toronto: University of Toronto Press.

Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Hambleton, R. K. (Ed.). (2000). Advances in performance assessment methodology [Special issue]. *Applied Psychological Measurement, 24*(4).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Harris, D., & Carr, J. (1996). *How to use standards in the classroom.* Alexandria, VA: Association for Supervision and Curriculum Development.

Hart, D. (1994). *Authentic assessment: A handbook for educators.* Menlo Park, CA: Addison-Wesley.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement, Issues and Practice, 18*(2), 5–17.

Levesque, K., Bradby, D., & Rossi, K. (1996). Using data for program improvement: How do we encourage schools to do it? *CenterFocus, 12.* Retrieved October 2, 2002 from http://ncrve.berkeley.edu/CenterFocus/CF12.html

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4–16.

Madaus, G., & Kellaghan, T. (1993). The British experience with "authentic" assessment. *Phi Delta Kappan, 74,* 458–469.

Madaus, G. F., & O'Dwyer, L. M. (1999). A short history of performance assessment: Lessons learned. *Phi Delta Kappan, 80,* 688–695.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Educational Policy Analysis Archives, 6*(13). Retrieved October 2, 2002 from http://epaa.asu.edu/epaa/v6n13.html

Mehrens, W. A., & Kaminski, J. (1989). Methods for improving standardized test scores: Fruitful, fruitless, or fraudulent? *Educational Measurement Issues and Practice, 8*(1), 14–22.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement, Issues and Practice, 14*(4), 5–8.

Meyer, R. H. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review, 16,* 283–301.

Miller, M. D., & Linn, R. L. (2000). Validation of performance-based assessments. *Applied Psychological Measurement, 24,* 367–378.

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

Moss, P. A., Beck, J. S., Ebbs, C., Matson, B., Muchmore, J., Steele, D., Taylor, C., & Herter, R. (1992). Portfolios, accountability, and an interpretive approach to validity. *Educational Measurement: Issues and Practice, 11*(3), 12–21.

Nagy, P. (1984). An examination of differences in high school graduation standards. *Canadian Journal of Education, 9,* 276–297.

Nagy, P. (1996). International comparisons of student achievement in mathematics and science: A Canadian perspective. *Canadian Journal of Education, 21,* 396–413.

Nagy, P., Traub, R. E., & Moore, S. (1999). A comparison of methods for portraying school demography using census data. *Alberta Journal of Educational Research, 45,* 35–50.

National Commission on Testing and Public Policy. (1990). *From gatekeeper to gateway: Transforming testing in America.* Boston: Boston College, National Commission on Testing and Public Policy.

Nichols, P. D., & Smith, P. L. (1998). Contextualizing the interpretation of reliability data. *Measurement: Issues and Practice, 17*(3), 24–36.

Ontario Ministry of Education and Training. (1997). *Putting students first: Ontario's plan for education reform* [leaflet]. Toronto: Queen's Printer.

Ontario Ministry of Education and Training. (2001, October 9). *School improvement teams created to improve reading skills* [news release]. http://www.edu.gov.on.ca/eng/document/nr/01.10/nr1009.html

Payne, K. J., & Biddle, B. J. (1999). Poor school funding, child poverty, and mathematics achievement. *Educational Researcher, 28*(6), 4–13.

Pennsylvania Department of Education. (n.d.). *2001 school performance funding: School-by-school results.* Retrieved October 2, 2002 from http://www.pde.state.pa.us/k12_initiatives/cwp/view.asp?a=173&Q=56493

Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice, 10*(4), 12–15.

Popham, W. J. (1999). Where large scale educational assessment is heading and why it shouldn't. *Educational Measurement: Issues and Practice, 18*(3), 13–17.

Robitaille, D. F., Taylor, A. R., & Orpwood, G. (1996). *The Third International Mathematics and Science Study, TIMSS-Canada Report: Volume 1, Grade 8.* Vancouver: University of British Columbia, Faculty of Education.

Ruiz-Primo, M. A., Baxter, G. P., & Shavelson, R. J. (1993). On the stability of performance assessments. *Journal of Educational Measurement, 30,* 41–53.

Serafini, F. (2001). Three paradigms of assessment: Measurement, procedure, and inquiry. *The Reading Teacher, 54,* 384–393.

Traub, R., Wolfe, R., Wolfe, C., Evans, P., & Russell, H. (1977). *Secondary-postsecondary interface project II: Nature of students.* Toronto: Ministry of Education and Ministry of Colleges and Universities.

Wohlstetter, P. (1991). Accountability mechanisms for state education reform: Some organization alternatives. *Educational Evaluation and Policy Analysis, 13,* 31–48.

Wolfe, R. G., Wiley, D., & Traub, R. (1999). *Psychometric perspectives for EQAO: Generalizability theory and applications* (EQAO Research Series No. 3). Toronto: Queen's Printer.

Worthen, B. R. (1993). Critical issues that will determine the future of alternative assessment. *Phi Delta Kappan, 74,* 444–454.