Broadening, Deepening, and Consolidating*

Alister Cumming
Ontario Institute for Studies in Education of the University of Toronto

Abstract

I encourage educators and researchers to continue, and to extend, three directions that are integral to the development of the field of language assessment: (a) to broaden the scope of inquiry and contexts that inform knowledge about language assessment; (b) to deepen the theoretical premises and philosophies of language assessment; and (c) to consolidate through systematic, critical reviews the information base of prior research on language assessment. To this end, I review various, notable studies conducted over the past decade that in my view exemplify these directions, suggesting topics or approaches that research and publications might productively pursue.

I express my hopes that three directions, prevalent in recent research on language assessment, continue to develop field of activity:

1. to broaden the scope of inquiry and contexts that inform knowledge about language assessment;
2. to deepen the theoretical premises and philosophies of language assessment; and
3. to consolidate through systematic, critical reviews the information base about prior research on language assessment.

Each of these directions has, over the past decades, made language assessment a more cognizant, scholarly, and responsible field than it was 30 or more years ago, when I found (in my initial university studies) that what was written about this topic tended merely (and uncritically) to prescribe practical steps to construct a grammar test (e.g., Lado, 1964) or to mark students’ essays or pronunciation errors (e.g., Rivers, 1968).

To this end, I have organized the present article to review recent publications that I believe have been especially valuable in refining and extending knowledge about language assessment. The publications I cite reflect my personal interests as well as my work in contexts where English is mostly the medium and subject of assessment. Though I have aimed for significant coverage internationally, I have not tried to be fully comprehensive nor definitive.
BROADENING

The scope of studies of language assessment can usefully be broadened in at least three ways. First, we need to know more about the practices and principles of assessment in ordinary contexts of teaching and learning. Second, we need to appreciate the full range of situations and populations around the world that engage in language assessment. Third, a broader range of methods of inquiry could productively be applied to studies of language assessment. Many publications have recently set the course for these directions, indicating what and how a broadened understanding of language assessment might entail.

Ordinary Assessment Practices

Understanding language assessment in ordinary educational practices is vital not only for the work of all teachers and program administrators but also for students whose language abilities are assessed and for those such as parents or employers who have a stake in the outcomes of language learning. The vast majority of books and articles published on language assessment have focused on formal tests, rather than routine assessment practices, presumably because research funding, major policy initiatives, and professional careers are associated with formal tests, not with ordinary teaching and learning. Widespread uses of the general term, “assessment,” rather than the narrower term “testing,” signal a shift in orientation to include a broad view of assessment in which inquiry into assessment practices in language classrooms and other settings feature prominently.

The graduate students I teach (who are usually experienced educators pursuing masters and doctoral degrees) certainly benefit from learning about formal language tests. And over the past decade, formal testing has assumed an increasingly greater role in their teaching practices and curriculum planning and reporting (such that many fundamental intersections now exist between performance on national or psychometric tests, implementing standard curricula, and helping students understand and shape their personal achievements). But their professional interests, and future applications of knowledge about language assessment, reside primarily in understanding better how to do and improve routine formative or achievement assessments. For example, the ordinary work of teachers requires them to monitor and report on student improvement, to link their lessons and classroom activities to system-wide standards, and to interact regularly with individual learners and other key stakeholders about these matters in an informed and responsible manner. From this
Broadening, Deepening, and Consolidating viewpoint, it is surprising how few systematic analyses there are of assessment practices in language classrooms, and thus how weak the knowledge base for pedagogical recommendations in this area actually are. Some notable exceptions include, for example, Rea-Dickins’ (2001) observational studies, Grierson’s (1995) survey of teachers and students, and my interviews with composition instructors in different countries (Cumming, 2001a), as well as studies of certain assessment functions where controversies over pedagogical practices have sparked research interest, such as instructors’ practices for responding to students’ writing (e.g., Ferris, 1997; Ferris & Roberts, 2001; Hyland & Hyland, 2001). But one is hard pressed to be able to cite many more publications that have tried systematically to describe language assessment practices in real classrooms as experienced by teachers, students, and program administrators. Moreover, few comprehensive frameworks exist to evaluate such practices critically. Such inquiry may be most valuable when assessment is viewed in respect to the whole context of a language curriculum, as Brindley (1998), Burns & Hood (1995), Darling-Hammond, Ancess and Falk (1995), Lynch and Davidson (1994), and TESOL (2001) have demonstrated in a range of different contexts. We need more studies that simply describe educators actually doing language assessment as well as critical analyses of the principles and variables these practices entail.

Various Situations of Language Assessment

Broadening the scope of situations in which research is conducted on language assessment is also necessary. Most published studies of language assessment have focused on young adults learning English in university contexts in English-dominant countries such as the United States, United Kingdom, Australia, New Zealand, or Canada. There is a certain amount of published research on the assessment of some widely-taught international languages--such as Arabic, French, German, Italian, Japanese, or Spanish--but most of this inquiry is situated in North America or Europe, and most of it in foreign language courses in universities or secondary schools. As a consequence, the knowledge base about language assessment is skewed toward these contexts and their particular interests, neglecting the majority of situations in schools, colleges, or informal adult education around the world in which language assessment happens. Not only do these local situations, and their unique requirements for language assessment, remain largely undocumented and unanalyzed, but a global perspective on their varying characteristics is lacking. It still astonishes me that despite considerable efforts in the early 1990s (documented in Cumming, 1996a),
demonstrations of the value of cross-cultural comparisons of language assessment (e.g., Bachman, Davidson, Ryan & Choi, 1995), and major efforts to benchmark language achievement across certain educational jurisdictions (e.g., North, 2000; Trim, 1998), nobody has been able, since John Carroll’s studies in the 1970s (e.g., Carroll, 1975), to organize an international comparative survey of second and foreign language achievement.

More than any topic, the graduate students, teachers, and program administrators I encounter express desires to know how to do language assessment with unique populations. Assessing young children’s language abilities is, for example, a topic they frequently ask about, but there remains little anyone can offer other beyond a few reports of exemplary innovations (e.g., Carpenter, Fujii & Katoaka, 1995; Johnstone, 2000), guidelines for excluding minority populations from standardized tests (e.g., Butler & Stevens, 2001; Solano-Flores & Trumbull, 2003), or pedagogically oriented advice (e.g., O’Malley & Peirce, 1996). That Douglas (2000) has been able to synthesize so many examples of language tests for specific academic, vocational, or professional purposes for adults underscores the need for more comprehensive documentation and careful analyses of such innovations in other contexts and among other populations. What of the knowledge that needs to be established on such topics as the assessment of language students with learning difficulties (e.g., Cummins, 1984)? Of aboriginal populations trying to retain their ancestral languages and cultural values (e.g., McGroarty, Beck & Butler, 1995)? Of teachers’ language proficiency (e.g., Elder, 2001)? Of determining the language proficiency required for specific professions (e.g., Epp, Stawychny, Bonham, & Cumming, 2002)? Of self-assessment (e.g., Ekbatani & Pierson, 2000)?

**Diverse Methods of Inquiry**

The repertoire of research methods established for the analyses of language tests now almost dictates that the properties of new assessment instruments need to be verified through analyses based on psychometrics (e.g., classical test theory, item response theory, or generalizability theory) and/or conventional statistical procedures (e.g., ANOVA, Regression analysis, Factor analysis, multi-faceted Rasch analyses, or Structural Equation Modeling) (cf. Davidson, 2000). Applications of these analytic methods are certainly to be encouraged to the extent that they provide evidence about the validity of language tests. But other, wholly different methods of inquiry also need to be adopted to achieve a broadened understanding of language assessment.
Serious consideration of the uses of language assessment requires adopting research methods that investigate people’s attitudes, beliefs, cultural values, and ways of interacting. For instance, studies using verbal reports have recently revealed many aspects of the thinking processes that examinees use to take language tests (e.g., Buck, 1992; Green, 1998) or that raters use to interpret and score the results of writing tests (e.g., Cumming, Kantor & Powers, 2002; Lumley, 2002; Milanovic, Saville & Shuhong, 1996). Likewise, rigorous discourse analyses have identified the types of conversational interaction that characterize oral proficiency interviews (e.g., Johnson, 2001; Lazaraton, 2002; Young & He, 1998). Such inquiry is indispensable for understanding why people perform the ways that they do in language assessments, and thus necessary for validation.

Many forms of qualitatively oriented inquiry could help to understand integral aspects of language assessment as well as to accommodate the interests and orientations of people, such as practicing educators, who regularly do language assessment. Policy analyses, for example, are fundamental to understand how language assessment figures in state policies and is acted on in education (e.g., Brindley, 1998; Cheng, 1997; McNamara, 1998; Slayter, 2003). Interviews, surveys and focus groups of key stakeholders in language assessment can illuminate their attitudes and beliefs, providing sources of criteria or verification for assessments (e.g., Cumming, Grant, Mulcahy-Ernt & Powers, 2004; Rosenfeld, Leung & Oltman, 2001). But why not probe more deeply into these matters in local contexts, for instance, through detailed case studies (e.g., Johns, 1991) or through phenomenological inquiry (e.g., Van Manen, 1990), as Tong (2000) did to study the conceptualizations of error that ESL writing instructors used in judging their students’ writing? Given the value of ethnography to describe cultural preferences and interactions, and the popularity of ethnographic methods in education and the social sciences generally, it is surprising how few ethnographies of language assessment practices there are (e.g., Dehyle, 1986; Losey, 1997). Likewise, I am at a loss to explain why I cannot think of a single application of narrative inquiry (e.g., Clandinin & Connelly, 2000; Egan, 1987) to examine teachers’ beliefs about language assessment (but see Bailey, 1998; Bell, 1997; Burns & Hood, 1995). Presumably, the epistemologies informing studies of language assessment have restricted the range of research methods adopted. As noted above, the emphasis on formal language tests has promoted analytic methods involving large sample sizes, appropriate to policy evaluation or investigating explicit hypotheses; but in striving for panoramic perspectives, such research
has tended to neglect the concerns for particular contexts, cases, and people that feature in qualitatively oriented inquiry and in pedagogically oriented decision making.

DEEPENING

Inherent in all methods of inquiry are the philosophical foundations or epistemologies that shape the questions that people pose for research. Broadening the scope of studies of language assessment in the ways suggested above may only have value if we also develop deeper theoretical foundations of what language assessment is. The most profound reconceptualization of language assessment this past decade has come in reference to Messick’s (1989) multi-faceted theory of the construct validity of tests (whose applications to language testing are documented, e.g., in Bachman, 1990; Chapelle, 1999; Cumming, 1996b; Kunnan, 1998, 2000). This single theory unifies the many kinds of validity evidence that can demonstrate that a test functions as it claims to. At the same time, the theory extends the range of situations that need to be investigated systematically to make this case, calling for investigations of the contexts in which a test is used in addition to the properties of the test itself.

Along with such theories of test development and use, we need theoretically informed and empirically substantiated models of language proficiency, learning, and development, as well as key sources of variability associated with them, to be able to make accurate and meaningful assessments and to explain their results (Bachman & Cohen, 1998). Although much work on language assessment since the early 1980s has usefully rallied around theoretical concepts of communicative competence, these concepts have proved to be difficult to operationalize and verify empirically in particular assessments (Bachman & Palmer, 1996; Harley, Allen, Cummins & Swain, 1990; Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000). New efforts at theory building are needed, for example, about language testing practices, assessment methods, and even the nature of language ability, social interaction, and learning. Hopefully, new ideas about these topics will feature in journals alongside reports of empirical data related to them. For instance, knowledge is gradually accumulating about particular aspects of second-language performance and development in specific domains. One line of development arises from the expansive scope of corpus-based studies (e.g., Biber, Conrad, Reppen, Byrd & Helt, 2002; Hawkey, 2003; Hinkel, 2002).
Another line of development concerns multiple perspectives on components of language ability. For example, for writing in second languages, models have emerged to account for such discrete components as composing processes (Sasaki, 2002), lexical knowledge and processing (Schoonen, van Gederen, de Glopper, Hulstijn, Snellings, Simis & Stevenson, 2002), grammatical concord (Mellow & Cumming, 1994), and textual indicators of long-term achievement (Cumming & Riazi, 2000; Grant & Ginther, 2000; Hinkel, 2002). Another decade of work in these directions may well produce empirical models able to guide language assessments explicitly. But theoretical conceptualizations are needed to make sense of empirical data beyond particular contexts. In the absence of such models, data, and theories, education systems around the world have been defining curriculum standards to prescribe expected levels of student achievement, in effect positing their own ad hoc models of language proficiency and development within their local curriculum contexts. Systematic inquiry into these initiatives is greatly needed (Cumming, 2001b; McKay, Coppari, Cumming, Graves, Lopriore & Short, 2001; North, 2000).

Sensing limitations in current conceptualizations of language assessment, scholars such as Peirce (1992), Valdes and Figueroa (1994), and Shohamy (2001) have urged educators to adopt a socially critical perspective, drawing on the ideas of Bourdieu (1991), Friere (1970), or Foucault (1979). A critical perspective aims to explain the dynamics of social power in the policies and practices of language assessment, for example, providing a basis to interpret ethical issues (e.g., Kunnan, 2000) or to judge the benefits of assessment policies for particular populations (e.g., Cumming, 1994). Indeed as Corson (1997) argued, all inquiry in applied linguistics could, if aligned with principles of critical realism, aim to serve emancipatory purposes.

But it seems to me that two levels of understanding have to be established prior to knowing what in language assessment might, or might not, constitute social justice or truly lead to emancipation. The first level is a theoretical framework capable of explaining relevant phenomena. Messick’s (1989) unified theory of construct validity has provided such a framework for formal language testing. But if we broaden our perspectives on language assessment to encompass all manners of educational and social policies and practices in situations around the world, then we must appeal to additional theories to explain language assessment in view of these other relevant phenomena. The theories of Bourdieu, Friere, or Foucault can help to explain, respectively, the social value of linguistic capital, the potential empowerment of disadvantaged groups, and the socio-historical basis of ideas such as
discipline. But the field of language assessment is still at a preliminary stage of establishing how to utilize and act on these concepts purposefully (and indeed is still grappling with the history and implications of statistical determinism, as outlined generally by Gould, 1996 or Hacking, 1990, and for language assessment, Davidson, 2004). Moreover, alternative, competing theories need to be examined. The possibilities here are infinite, but as examples I can cite just two philosophers whose ideas are commonly applied in educational research, those of John Dewey (e.g., 1988; Prawat, 1998) to explain assessment in view of teaching and curricula or of Jurgen Habermas (e.g., 1984; Ewert, 1991) to explain assessment in view of social policies.

The second level of understanding concerns defining core concepts. In some respects, this requires scholarly reflection and analysis (as in Alderson, Clapham & Wall, 1995; Bachman, 1990; Davies, 1990) as well as collaboration to compile tools like dictionaries of key terms (e.g., Davies, Brown, Elder, Hill, Lumley & McNamara, 1999) or to survey jurisdictions to establish common levels of professional standards (e.g., Alderson, Davidson, Douglas, Huhta, Turner & Wylie, 1995). But in another respect, each concept integral to language assessment needs to be examined rigorously through empirical scrutiny. This approach has featured in several especially valuable studies of language assessment, which have critically examined such taken-for-granted ideas as washback (Alderson & Hamp-Lyons, 1996; Bailey, 1996; Burrows, 2001; Wall & Alderson, 1996), oral fluency (Fulcher, 1996), or grammaticality judgments (Birdsong, 1989). Ultimately, these levels of core concepts, theoretical frameworks, and empirical data need to be brought together to establish a working professional consensus. I think, as an analogy, of Hornberger’s (1989, 2003) model of continua to explain individual and societal variability in reading and writing among bilinguals. Like studies of biliteracy, the field of language assessment has been entering a similar phase of elaborating explanatory models, accounting for ordinary as well as diverse phenomena, and refining the sense of core constructs. There is much to look forward to in this regards in the future, and much for active researchers and thinkers to contribute.

CONSOLIDATING

To know what or how to broaden and deepen studies of language assessment requires critical awareness of the relevant knowledge that is already established and that is worthy of
further consideration. Consolidating such information will, I hope, be a further goal of future inquiry (including, I would like to think, the present article). Critical review articles have played an important role in defining salient issues, new directions, and topics worthy of further examination in language assessment. Some have highlighted notable trends from a broad, comprehensive perspective (e.g., Alderson & Benerjee, 2001; Bachman, 2000; Kunnan, 1999), whereas others have usefully demarcated topics of specific interest, such as computer-adaptive language testing (e.g., Chalhoub-Deville & Deville, 1999) or concerns in analyses of language assessment policies (McNamara, 1998). Another example is the set of reviews of theories and empirical research in the framework reports prepared to guide the development of a new TOEFL; these reviews consolidate current knowledge about the assessment of English proficiency in academic settings in ways that not only express the current state of informed opinion but are also sure to shape future thinking about the testing of listening, speaking, writing, and reading abilities (Bejar, Douglas, Jamieson, Nissan & Turner, 2000; Butler, Eignor, Jones, McNamara, & Suomi, 2000; Cumming, Kantor, Powers, Santos & Taylor, 2000; Enright, Grabe, Koda, Mosenthal, Mulcahy-Ernt & Schedl, 2000). At the same time, reports are emerging on the major developments taking place at many levels across the European Community, involving languages other than English, ranging from self-assessments (e.g., Dialang) to the implementation of comprehensive curriculum policies (e.g., Trim, 1998). Compiling, categorizing, and annotating bibliographies likewise help to consolidate and make accessible information published in disparate sources (e.g., Banerjee, Clapham, Clapham & Wall, 1999).

In addition to thematically or conceptually organized review articles, along with the types of consolidation through theory building I describe above, two particular types of studies are especially valuable. One is meta-analysis and other forms of research synthesis (Cooper & Hedges, 1994; Wolf, 1986). The few systematic meta-analyses that have been conducted in relation to language assessment (e.g., Dochy, Segers, & Buehl, 1999; Norris & Ortega, 2000; Ross, 1998) contribute so importantly to consolidating current knowledge, and provide so many implications for assessment practices, that one can only hope to read and learn from more such synthetic studies along the lines of articles that regularly appear in the Review of Educational Research. The other type of article would be chronicles and analyses of the history of particular aspects of language assessment. Publications such as Barnwell (1996), Milanovic and Weir (2003), and Spolsky (1995) stand out so uniquely in their historical perspectives on this field that they call for others to embark seriously on historical
studies of how and why language assessment has developed the way that it has. Such inquiry might be especially valuable in a context such as Taiwan, where a new national test of English has recently been developed, building on past ideas and practices that are worth documenting systematically.

**SUMMARY**

I appreciate that, on the one hand, I have charted out a "wish list" that may be idiosyncratic in expressing my personal interests as well as challenging for others to fulfill. But on the other hand, I presume that the directions I have outlined conform to many of the reasons for the present conference to develop awareness of international trends in language assessment. We all want to see more studies of language assessment that broaden, deepen, and consolidate our knowledge and abilities.

**ACKNOWLEDGEMENTS**

This article originally appeared in *Language Assessment Quarterly, 1*, 1, 5-18. Permission to reprint from Lawrence Erlbaum Associates, Inc. is gratefully acknowledged.

**REFERENCES**


