BONNY NORTON

ACCOUNTABILITY IN LANGUAGE ASSESSMENT

In many language assessment projects, stakeholders include test takers, teachers, test developers, administrators, community agencies, public officials, and researchers. Increasingly, calls for accountability in language assessment have focused on the consequences of assessment practices for test takers, who have hitherto been relatively powerless stakeholders in the field of language assessment. The recognition that language assessment practices should be accountable to test takers is indicative of the broader trend towards democratization of educational assessment in general. Such concerns are equally important in debates on ethics in language testing and language testing standards (see the reviews by Hamp-Lyons, and Davidson, Turner and Huhta in this volume.)

EARLY DEVELOPMENTS

In his article, 'The measurement of communicative competence', Canale (1988) argued that what he called 'the naturalistic-ethical approach' should complement the three trends in language testing outlined by Spolsky (1978). These trends were the traditional approach, focusing on language testing as an art; the psychometric-structuralist approach, focusing on language testing as a science; and the integrative-sociolinguistic approach, focusing on language testing as both art and science. Canale's main argument was that any information, once gathered, can be used unethically, and that language testers are to some extent accountable for ensuring that the information they gather is used for ethical purposes. Drawing on Messick (1981), he argued that language testers need to consider the social consequences that the use of a given test may have and that adequate training in testing and test use is required if language test developers and test users are to be professionally accountable.

In advocating a fourth approach to language assessment, Canale was subscribing to the views of those scholars who were arguing for greater accountability to test takers. These scholars included Cummins, Cohen, Deyhle, Fillmore, and Swain. Cummins (1984) argued that there was confusion in the language assessment of students in bilingual programmes because of the failure to develop an adequate theoretical framework for relating language proficiency to academic achievement. Cohen (1984) conducted research on second language test-taking strategies and concluded that there should be a better fit between the expectations of test developers

C. Clapham and D. Corson (eds), Encyclopedia of Language and Education, Volume 7: Language Testing and Assessment, 313–322.
© 1997 Kluwer Academic Publishers. Printed in the Netherlands.

on the one hand and test takers on the other. Deyhle's (1986) research with Navajo students demonstrated that test taking itself is a cultural activity and that Navajo and Anglo students had very different conceptions of testing. Fillmore (1982) found that the texts used in reading comprehension tests for young children were highly problematic, requiring an "uncommon degree of tolerance and co-operation" (p. 251) on the part of the test taker. Swain (1985) argued that test developers should "bias for best" in the test development process in order to give test candidates the greatest opportunity to demonstrate their knowledge of the target language. All this research represents growing recognition that test takers come from heterogeneous, culturally diverse backgrounds that must be taken seriously in the assessment process.

MAJOR CONTRIBUTIONS

The drive towards greater accountability to test takers has been gathering momentum (see Kunnan, 1996). Language testers have shown increasing interest in the investments that test takers have in language tests and have raised concerns about the limitations of existing instruments. The practice and theory of accountability in language testing is being explored in different countries and with different populations of test takers. Sometimes, the focus of the work is on language testing per se; sometimes, language testing is integrated into broader initiatives for greater accountability in educational assessment.

Basing her research on Nigeria, Parry (1994) examines material from the English examination of the West African School Certificate (WASC). She demonstrates that there are subtle contradictions between what is stated in the reading passages and what is implied in the tasks. In Israel, Shohamy (1993a) conducted research on Arabic as a second language, English as a foreign language, and Hebrew as a first language, and concluded that the use of tests to solve educational problems is a simplistic approach to a complex problem. In the United States, research has illustrated the disjuncture between test taker identities and investments on the one hand, and large-scale standardized language tests on the other. Lowenberg's (1993) research on the Test of English for International Communication (TOEIC) calls into question the assumptions on which the test is based, arguing, in particular, that what constitutes standard English is open to international debate. Peirce's (1992) research on the Test of English as a Foreign Language (TOEFL) reading test highlights the unequal relationship between test makers and test takers and focuses on the need for greater accountability to test takers. Raimes (1990) raises concerns about the validity of the TOEFL Test of Written English and the washback effects the test will have (see Cumming's review in this volume). In Canada, Elson (1992) examines English as a Second Language (ESL) proficiency testing in the

admissions process at Canadian universities. He argues that universities use language proficiency tests in such a way as to avoid taking responsibility for the educational needs of students from diverse backgrounds. With reference to South Africa, Peirce & Stein (1995) examine an alternative university admissions test and argue that there is a discrepancy between the test format and its purpose. In England, Alderson & Buck (1993) conducted a survey of British examination boards that offer English language tests and questioned whether they met high standards in educational measurement. Matthews (1990), likewise, has raised concerns about the assessment criteria of international examinations of English as a foreign language, particularly with respect to the assessment of productive skills.

Because concerns have been raised about the extent to which large-scale language assessment is accountable to test takers, there have been concerted efforts at more local levels to achieve greater accountability. Cohen (1994) describes his research on the innovative use of summary tasks to assess proficiency in reading English in a Brazilian university context. In Australia, the Bandscales project (McKay, 1995; Moore, 1996) places ESL development in the context of the school and its curriculum. By directing ESL programs and teachers to 'across the curriculum' concerns, the Bandscales developers seek to be accountable to language learners. (For more about this work, see McKay's review in this volume.) Kalantzis, Cope & Slade (1989), alternatively, argue that Halliday's systemic-functional linguistics can provide a comprehensive framework for assessment because it overcomes the competence/performance dichotomy that characterizes much linguistic theory. With reference to Zimbabwe, Allen (1994) describes the weakness of the Zimbabwe Junior Certificate exam and outlines principles on which a new test should be based. In Japan, Ingulsrud (1994) examines the reading assessment component of the Joint First Stage Achievement Test (JFSAT) that governs entrance into Japanese universities. He demonstrates how test developers are required to defend their practices in a public forum. In South Africa, Duncan (1995) draws on both quantitative and qualitative assessment to demonstrate the effectiveness of an innovative project to enhance both mother-tongue and English language proficiency, while Yeld & Haeck (1993) describe how assessment instruments can be used to promote the access of disadvantaged students to tertiary education.

In the United States, Lacelle-Peterson & Rivera (1994) argue that educational reform that serves anglophone students will not necessarily benefit English language learners; to address this problem, they provide a useful framework for the equitable assessment of English language learners. In an edited volume, Holland, Bloome & Solsken (1994) draw together a number of innovative assessment projects addressing the language and literacy skills of young school children. The articles represent three broad theoretical and disciplinary perspectives that are gaining wider acceptance

in educational practice: anthropological, sociopsycholinguistic, and reader response. In Canada, Cumming (1994) reviews the functions of language assessment for recent immigrants to Canada and argues that assessment instruments should meet the important criterion of facilitating the participation of immigrants into Canadian society. Larter & Donnelly (1993) describe the development of the Toronto Benchmark Program, arguing that it provides a framework that is accountable to students, parents, and teachers. Wesche (1987) describes the potential of a performance-based assessment instrument to be used for post-secondary admission in Ontario. In the United Kingdom, Holland & Street (1994) describe the methods of assessing literacy skills developed by the Adult Literacy Basic Skills Unit, focusing in particular on the 'Progress Profile'. All of these initiatives represent innovative responses to the question of how test makers can be more accountable to test takers.

WORK IN PROGRESS

Currently there is much work in progress on the development of codes of practice and benchmark standards in language testing. Some of this work is informed by the development of the American Psychological Association (APA) Standards and the Code of Fair Testing Practices in Education (see Stansfield, 1993). The Association of Language Testers of Europe (ALTE) has developed a Code of Practice to make explicit the standards they aim to meet and the responsibilities they have undertaken (ALTE, 1994). ALTE represents language testing organizations in France, Spain, the Netherlands, Portugal, Denmark, Italy, Germany and the United Kingdom. The Code of Practice distinguishes between the interests of examination developers, examination users, and examination takers, and its central undertaking is "to safeguard the rights of examination takers" (ALTE, 1994, p. 4). In Canada, the federal government has sponsored the development of a document that provides a common framework for describing the language skills of adult clients across the country (Citizenship and Immigration Canada, 1996). This document addresses the needs of adult learners of ESL as well as adult literacy learners. Assessment instruments to be used for placement purposes have been developed in accordance with this document (Peirce & Stewart, 1997). In the international TESOL (Teachers of English to Speakers of Other Languages) organization, a working group has been established to review key issues in assessing the language development of English language learners. This development is an outgrowth of the ESL Standards Project (Katz & Short, 1996). Likewise, the International Language Testing Association (ILTA) is in the process of developing a Code of Practice for Language Testers internationally. (For more about such codes of practice, see the review by Davidson, Turner and Huhta in this volume.)

There are an increasing number of conferences that are focusing on accountability to test takers. In April 1992, a symposium on 'Testing English Across Cultures' was commissioned as part of the 'World Englishes Today' conference held in Urbana, Illinois, USA. As Davidson (1993, p. 114) notes, this conference raised questions about the variety of English that is promoted in international tests of English, and the "consequent irrelevance of the concept of 'native speaker' ". In Turfloop, South Africa, in October 1994, the South African Association for Academic Development (SAAAD) held a conference with the theme, 'Accountability in Testing'. The central question addressed at the conference was as follows: How can tests be used innovatively to increase the access of historically disadvantaged English language learners to tertiary education in South Africa? At the second International Conference on Evaluation, held in Toronto, Canada, in October 1995, two plenary speakers, Linda Darling-Hammond and Caroline Gipps, addressed accountability in assessment. At the annual meeting of the Language Testing Research Colloquium (LTRC) in Long Beach, California, in March 1995, the theme of the conference was 'Validity and Equity Issues in Language Testing'. Only two years later, in 1997, the theme of this same conference, held in Orlando, Florida was 'Fairness in Language Testing'. (For more about other presentations at international conferences, see Hamp-Lyons' review in this volume.) It is anticipated that all this work in progress will generate further research and publications on accountability in language assessment.

PROBLEMS AND DIFFICULTIES

Even as interest in accountability towards individual test takers has increased, there has also been a growing concern in the field of both educational assessment and language assessment that assessment practices should address system-wide accountability. Schools, colleges and universities are under pressure to inform the public about what they are teaching and how effective they are (see Darling-Hammond, 1994; Earl, 1995; Froese, 1997; Gipps, 1994). (For more about accountability and league tables see the review by Rea-Dickins and Rixon in this volume.) In this climate, there is tension between formative and summative assessment – what Gipps (1994, p. 12) calls a "paradigm clash". As Brindley (1995) argues:

> Teachers thus are now finding themselves under pressure from two different directions. On the one hand, they need to carry out detailed individual assessments at the individual level for purposes of diagnosis and feedback to learners, a role they are prepared to embrace because of the obvious effects on instruction (Broadfoot, 1992). However, at the same time, they are

increasingly being called on to report learners' progress against national standards in order to meet accountability requirements.

The tension between these orientations is exacerbated by the fact that, while much research questions the validity of large-scale, system-wide assessment (see the section on 'Major Contributions' above), there is little research to support the validity of more learner-oriented assessment, such as performance assessment and portfolio assessment (Hamp-Lyons, 1996; Reardon, Scott & Verre, 1994). In the field of language assessment, McNamara (1995) identifies a number of problematic features of performance assessment in the main models proposed in the second language assessment context, particularly with regard to the interactional aspect of performance. Hill & Parry (1994, p. 146) point to some of the problems with alternative forms of assessment, focusing on the extent to which they meet reliability criteria. Furthermore, as Darling-Hammond (1994) argues, alternative assessment methods, that appear more accountable to learners, are not necessarily equitable. She indicates that assessment reform is sometimes used as a means for external control of schools and stems from a distrust of teachers. She suggests that the equitable use of alternative assessments depends not only on the design of the assessments themselves, but also on the extent to which teachers are an integral part of the reform process. However, as Gearhart & Herman (1995) demonstrate, attempts to integrate learner-oriented, performance assessment in large-scale, system-wide assessment pose many challenges for all stakeholders.

FUTURE DIRECTIONS

The challenges for the future must be understood in relation to the struggles of the past and the possibilities of the present. Firstly, there is a need to address the tension between accountability to individuals on the one hand and accountability to systems on the other. In this regard, the work of Fulcher & Bamford (1997) raises important questions. Fulcher and Bamford examined standards in language testing in the context of the legal framework of the United States of America (USA) and the United Kingdom (UK). They argue that while the threat of litigation has generated much research on the reliability and validity of educational assessment instruments in the USA, the same does not apply to the UK. They conclude that it may only be a question of time before test takers in the UK (and by extension other nations) seek legal means to ensure that language testers are accountable to test takers. Current debates in the USA on the 'opportunity to learn' (Guiton & Oakes, 1995) may become central in the quest for accountable language assessment. Over time, innovative language testing theory and practice may engender a complementary rather than adversarial relationship between disparate stakeholders in language testing practices.

Secondly, there may be increased research on the washback effects of

testing on teaching and learning. Xiaoju's (1990) research on the Matriculation English Test (MET), a new English matriculation test in China, for example, indicates that language testing can encourage innovative teaching practices. Research on Alderson & Wall's (1993) 'washback hypotheses' will make an important contribution to debates on accountability. An interesting avenue for research will be an investigation of ways in which teaching can inform testing. In this spirit, as Lacelle-Peterson & Rivera (1994), Peirce (1992) and Shohamy (1993b) have argued, if language testers are to be accountable to test takers, it will be necessary for them to enter into a dialogue with a broad range of stakeholders so that teaching and learning can be enhanced. (For more about washback, see Wall's review in this volume.)

Thirdly, language testers may become increasingly interested in the possibilities of computer adaptive testing (CAT) in attempts to be more accountable to test takers (see Jones, 1994; Tung, 1986). As Tung (1986) argues, not only does CAT take less time to administer than many other forms of assessment, but it may produce desirable 'affective effects' on test takers, who will find that while the test items are always challenging for them, they will seldom be beyond their capability. (For more about CAT, see Gruba and Corbel's review in this volume.)

Whatever trends in language assessment emerge in the next millennium, accountability in language assessment will remain a central priority for all stakeholders.

University of British Columbia, Canada

REFERENCES

Alderson, J.C. & Buck, G.: 1993, 'Standards in testing: A study of the practice of UK examination boards in EFL/ESL testing', *Language Testing* 10(1), 1–26.

Alderson, J.C. & Wall, D.: 1993, 'Does washback exist?', *Applied Linguistics* 14(2), 115–129.

Allen, K.: 1994, 'English education in Zimbabwe: Testing communicative competence', in C. Hill & K. Parry (eds.), *From Testing to Assessment: English as an International Language*, Longman, London.

Association of Language Testers of Europe: 1994, *The ALTE Code of Practice: The Code of Practice for the Association of Language Testers in Europe*, Version 1, January, 1994.

Brindley, G.: 1995, *Assessment and Reporting in Language Learning Programs: Purposes, Problems, and Pitfalls*, Plenary given at the International conference on Testing and Evaluation in Second Language Education, Hong Kong University of Science and Technology, 21–24 June, 1995.

Broadfoot, P.: 1992, *A Question of Quality: The Changing Role of Assessment in Education*, ACSA Workshop Report No. 4, Australian Curriculum Studies Association, Canberra.

Canale, M.: 1988, 'The measurement of communicative competence', *Annual Review of Applied Linguistics* 1987, 8, 67–84.

Citizenship and Immigration Canada: 1996, *Canadian Language Benchmarks: English as a Second Language for Adults/English as a Second Language for Literacy Learners*, Working Document, Minister of Supply and Services, Ottawa, Canada.

Cohen, A.: 1984, 'On taking language tests: What the students report', *Language Testing* 1(1), 70–81.

Cohen, A.: 1994, 'English for Academic Purposes in Brazil: The use of summary tasks', in C. Hill, & K. Parry (eds.), *From Testing to Assessment: English as an International Language*, Longman, London.

Cumming, A.: 1994, 'Does language assessment facilitate recent immigrants' participation in Canadian Society? *TESL Canada Journal* 2(2), 117–133.

Cummins, J.: 1984, 'Wanted: A theoretical framework for relating language proficiency to academic achievement among bilingual students', in C. Rivera (ed.), *Language Proficiency and Academic Achievement*, Multilingual Matters, Clevedon, Avon.

Darling-Hammond, L.: 1994, 'Performance-based assessment and educational equity', *Harvard Educational Review* 64, 1.

Davidson, F.: 1993, 'Testing English across cultures: Summary and comments', *World Englishes* 12(1), 113–125.

Deyhle, D.: 1986, 'Success and failure: A micro-ethnographic comparison of Navajo and Anglo students' perceptions of testing', *Curriculum Inquiry* 16(4), 365–389.

Duncan, K.: 1995, 'The role of testing in the evaluation of a primary education project: The case of Molteno', in P. Rea-Dickins & A.F.L. Lwaitama (eds), *Evaluation for Development in English Language Teaching*, ELT documents, Macmillan Publishers Limited, London and Basingstoke, 107–116.

Earl, L.M.: 1995, 'Assessment and accountability in education in Ontario', *Canadian Journal of Education* 20(1), 45–55.

Elson, N.: 1992, 'The failure of tests: Language tests and post-secondary admissions of ESL students', in B. Burnaby & A. Cumming (eds), *Socio-Political Aspects of ESL in Canada*, OISE Press, Toronto.

Fillmore, C.: 1982, 'Ideal readers and real readers', in D. Tannen (ed.), *Analyzing Discourse: Text and Talk. Georgetown University Round Table*, 1981. Georgetown University Press, Washington, DC.

Froese, V.: 1997, 'National assessment the Canadian way', *Reading Today*, February/March 1997, 26.

Fulcher, G. & Bamford, R.: 1997, 'I didn't get the grade I need. Where's my solicitor?', *System*, 437–448.

Gearhart, M. & Herman, J.: 1995, *Portfolio Assessment: Whose Work Is It? Issues in the Use of Classroom Assignments for Accountability*, UCLA Center for the Study of Evaluation, Los Angeles, CA.

Gipps, C.V.: 1994, *Beyond Testing: Towards a Theory of Educational Assessment*, Falmer Press, Washington, D.C.

Guiton, G. & Oakes, J.: 1995, 'Opportunity to learn and conceptions of educational equality', *Educational Evaluation and Policy Analysis* 17(3), 323–336.

Hamp-Lyons, L.: 1996, 'Applying ethical standards to portfolio assessment of writing in English as second language', in M. Milanovich & N. Saville (eds.), *Performance Testing, Cognition and Assessment: Selected Papers from the 15th Language Testing Research Colloquium*, Cambridge University Press, Cambridge.

Hill, C. & Parry, K. (eds.): 1994, *From Testing to Assessment: English as an International Language*, Longman, London.

Holland, K., Bloome, D. & Solsken, J.: 1994, *Alternative Perspectives in Assessing Children's Language and Literacy*, Ablex, Norwood, N.J.

Holland, D. & Street, B.: 1994, 'Assessing adult literacy in the United Kingdom: The progress profile', in C. Hill & K. Parry (eds.), *From Testing to Assessment: English as an International Language*, Longman, London.

Ingulsrud, J.E.: 1994, 'An entrance test to Japanese universities: Social and historical context', in C. Hill & K. Parry (eds.), *From Testing to Assessment: English as an International Language*, Longman, London.

Jones, N.: 1994, *Adaptive Testing and Adaptive Learning*, Paper given at the Language Testing Forum, Cambridge, England, December, 1994.

Kalantzis, M., Cope, B. & Slade, D.: 1989, *Minority Language and Dominant Culture: Issues of Education*, Assessment and Social Equity, Falmer Press, London.

Katz, A. & Short, D.: 1996, 'ESL standards and the TESOL assessment guidelines project', *TESOL Matters* 6(2), 1 and 14.

Kunnan, A.J.: 1996, *Connecting Fairness with Validation in Language Testing*. Paper presented at the 18th Language Testing Research Colloquium, Tampere, Finland, August, 1995.

Lacelle-Peterson, M.W. & Rivera, C.: 1994, 'Is it real for all kids? A framework for equitable assessment policies for English language learners', *Harvard Educational Review* 64, 1.

Larter, S. & Donnelly, J.: 1993, 'Demystifying the goals of education: Toronto's benchmark program', *Orbit* 24(2), 22–28.

Lowenberg, P.: 1993, 'Issues of validity in tests of English', *World Englishes* 12(1), 95–106.

Matthews, M.: 1990, 'The measurement of productive skills: Doubts concerning the assessment criteria of certain public examinations, *ELT Journal* 44(2), 117–121.

McKay, P.: 1995, 'Developing ESL proficiency descriptions for the school context: The NLLIA ESL bandscales', in G. Brindley (ed.), *Language Assessment in Action*, National Centre for English Language Teaching and Research, Macquarie University, Sydney.

McNamara, T.F.: 995, 'Modelling performance: Opening Pandora's box', *Applied Linguistics* 16(2), 159–179.

Messick, S.: 1981, 'Evidence and ethics in the evaluation of tests', *Educational Researcher* 10(9), 9–20.

Moore, H.: 1996, 'Telling what is real: Competing views in assessing ESL development', *Linguistics and Education* 8(2), 189–228.

Parry, K.: 1994, 'The test and the text: Readers in a Nigerian secondary school', in C. Hill & K. Parry (eds.), *From Testing to Assessment: English as an International Language*, Longman, London.

Peirce, B.N.: 1992, 'Demystifying the TOEFL reading test, *TESOL Quarterly* 26(4), 665–689.

Peirce, B.N. & Stein, P.: 1995, 'Why the monkeys passage bombed: Tests, genres and teaching', *Harvard Educational Review* 65(1), 50–65.

Peirce, B.N. & Stewart, G.: 1997, 'The development of the Canadian language benchmarks assessment', *TESL Canada Journal* (in press).

Raimes, A.: 1990, 'The TOEFL Test of Written English: Causes for concern', *TESOL Quarterly* 24, 427–442.

Reardon, S., Scott, K., & Verre, J.: 1994, 'Symposium: Equity in educational assessment', *Harvard Educational Review* 64(1), 1–4.

Shohamy, E.: 1993a, *The Power of Tests: The Impact of Language Tests on Teaching and Learning*, National Foreign Language Center Occasional Paper, Washington, D.C.

Shohamy, E.: 1993b, 'The exercise of power and control in the rhetorics of testing', in A. Huhta, K. Sajavaara & S. Takalo (eds.), *Language Testing: New Openings*, Institute for Educational Research, Jyväskylä, Finland.

Spolsky, B.: 1978, 'Introduction: Linguists and language testers', in B. Spolsky (ed.), *Advances in Language Testing Research: Approaches to Language Testing*, Volume 2, Center for Applied Linguistics, Washington, D.C.

Stansfield, C: 1993, 'Ethics, standards, and professionalism in language testing', *Issues in Applied Linguistics* 4(2), 189–206.

Swain, M.: 1985, 'Large-scale communicative language testing: A case study', in S. Savignon & M. Burns (eds.), *Initiatives in Communicative Language Teaching*, Addison-Wesley, Reading, MA.

Tung, P.: 1986, 'Computerized adaptive testing: Implications for language test developers', in C. Stansfield (ed.), *Technology and Language Testing*, TESOL, Washington, D.C.

Wesche, M.: 1987, 'Second language performance testing: The Ontario Test of ESL as an example', *Language Testing* 4(1), 28–47.

Xiaoju, L.: 1990, 'How powerful can a language test be? The MET in China', *Journal of Multilingual and Multicultural Development* 11(5), 393–404.

Yeld, N. & Haeck, W.: 1993, 'Educational histories and academic potential: Can tests deliver?', in S. Angelil-Carter (ed.), *Language in Academic Development at U.C.T.*, unpublished manuscript.