*ReCALL* **12** (1): 93–104. Printed in the United Kingdom

© 2000 Cambridge University Press

# *Can computerised testing be authentic?*

MICHEL LAURIER
*University of Montreal, Montreal, Canada*
(*e-mail: laurierm@SCEDU.Umontreal.ca*)

**Abstract**

The concept of authenticity first appeared with the development of the communicative approach. More recently, in the field of educational measurement, authentic assessment methods have been proposed. Although adaptive testing seems to be the most important application of computers in language assessment, these tests are usually not authentic. Since many real world tasks are accomplished with computers, these may be used for authentic direct testing. Computers may be also used in semi-direct testing as a way to enhance the context. Finally in authentic assessment, computers may be used as a tool to process the data when the learners use them to organise their portfolio. Using the computer, test developers can also create better authentic tests.

## 1 Introduction

In second language learning and teaching, we often refer to the concept of authenticity. Computer assisted language learning raises the issue of authenticity in a particular way since the learner is placed in a context that is different from a real communication situation. This assumption has to be discussed but it is certainly true that in most cases, interaction mediated with a computer is different from a face-to-face interaction between two human beings. Even when learning focuses on writing skills, the environment is often rather artificial. In other respects, ensuring authenticity in testing or assessment situations is a major issue. Recent developments in the field of educational measurement show that creating authentic situations that allow inferences about learners' competence is not always easy. In this paper, I intend to clarify the concept of authenticity from a language learning perspective and from an educational measurement perspective. Then I will try to answer the initial question: Can computerised testing be authentic?

## 2 The concept of authenticity

The concept of authenticity in language learning and teaching appeared with the development of the communicative approach at the beginning of the 1980s. At that time, Canale (1983: 2–3) reformulated Morrow's (1979) concern with the nature of communication for testing. In this reformulation, he presented seven features of communication that had to be taken into consideration for testing. Communication...

- "is interaction-based": language is normally acquired and used in social interaction;

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

- "involves unpredictability and creativity";
- is constrained by the "discourse and sociocultural context";
- is constrained by "psychological and other conditions": memory, noise...;
- always has at least one purpose;
- "involves authentic as opposed to textbook-contrived language";
- is judged as successful "on the basis of actual outcomes".

Many authors emphasised the point advocating that the use of authentic documents in the language classroom should be favoured. In this perspective, material that had been created specifically for the language class or even adapted material was deemed as unnecessary. Of course, as Widdowson (1979) pointed out, the authenticity of a document is questionable when it is used outside its normal context, in a classroom. To avoid confusion and illusion about the authenticity of material that could be used for pedagogical purposes, Widdowson made a distinction between "genuine" and "authentic" documents. Later, pedagogical experience of authentic material in the classroom and the theoretical reflection triggered by Krashen's comprehensible input hypothesis (Krashen, 1982) have somewhat tempered the enthusiasm for authentic documents.

The application of the communicative approach added another dimension to authenticity, which is the design of syllabuses in terms of language tasks that are normally performed in a natural setting. One can certainly imagine academic or professional situations which a learner is likely to encounter in real life in second language learning. However, should we say that a conversation in French about a maths problem between two students learning a second language in an immersion programme in Vancouver is less authentic than the simulation of a meal in a high-class Parisian restaurant? In many situations, the classroom setting is perhaps more authentic than any replication of a situation involving native speakers. In this regard, we must not lose sight of an important distinction made by Stern (1981) who defined two types of authenticity. On the one hand, content authenticity refers, first, to the type of language that should be used in a given situation and the avoidance of unnatural language and, secondly, to the use of real documents in the classroom such as newspaper articles or taped live conversations. On the other hand, learning authenticity refers first to the student's commitment to his or her own learning and the motivation to use the language in real life and secondly to the type of interactions which should be spontaneous and meaningful instead of contrived and artificial.

## 3 The authentic assessment movement

Although not entirely new for language testers, the authentic assessment trend is very important as it seriously challenges traditional beliefs in testing and assessment and it is now changing our educational values. Wiggins (1993), who acted as the spokesman of the authentic assessment movement, describes authentic assessment with two main characteristics:

- The task must be contextualised – it is not sufficient to ask students to talk about their hobby even though they may have to do it in real life. The description of the task must also specify the context: who is the interlocutor? under which circum-

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

stances? what is the purpose of presenting your hobby? and so on. The task must include all the information that may be relevant to replicate a performance that is valued in the real world outside the school.

- The task must be complex – assessment tasks which focus on specific isolated objectives that call for the application of one single strategy, one ability, one element of competence are not useful. The task must involve the interaction of different types of knowledge that are not necessarily identical from one learner to the other and may even lead to different but worthwhile solutions.

It is interesting to notice that Wiggins (1989), in an earlier article, had quoted the ACTFL oral interview as an example of an authentic test because the oral interview replicates a situation that the learner may encounter in real life. However, what Wiggins understated is that interviewers must be trained to elicit performances that can be rated according to standards that are not taken into account in a non-test situation (Wesche, 1992).

Wiggins claims that psychometric approaches to student assessment have been detrimental to learning because they induce teaching and learning practices that do not allow the integration of knowledge and the development of problem solving strategies. In authentic assessment, the assumption that contextualised and complex tasks necessarily bring into play abilities that we ought to measure, means shifting the priority given to validity at the expense of reliability. The concept of authentic assessment is sometimes used synonymously with performance assessment, an approach that is increasingly recognised as a viable alternative to classical psychometric approaches. Performance assessment is intended to narrow the gap between the observed performance, what the student does, and the interpretation, the inference that is made in order to produce a judgment about the competence that is required to obtain such a performance. Since the inference from the observed performance to conclusions about the competence seems rather direct, many proponents of performance assessment consider a task presented within this framework as necessarily valid (Messick, 1994). However, Wiggins (1998: 21) warns us that authentic assessment is not synonymous with performance assessment when he replies to Terwilliger's (1997) measurement concerns about new forms of assessment: "Many so-called performance tasks or constructed response items do not replicate the conditions under which adults are challenged in context". And he further adds: "The point is not to measure but to improve". In other words, in Wiggins' perspective authentic assessment is valid not only because of the correspondence with real life conditions, but also because it improves learning as students may eventually enhance their capability to face real world situations. From this point of view, authentic assessment takes into account the concerns about the impact of assessment as raised by some researchers almost ten years ago (Fredericksen and Collins 1989; Messick 1989). It meets the requirements of consequential validity as described by Messick (1989, 1994) who advocates that making sure our testing procedures adequately measure the skills that are supposed to be measured is not sufficient. According to Messick, testing procedures should also have a positive effect on learning.

The importance of assessment procedures impacting on learning is illustrated by the use of portfolios which has been proposed, within the authentic assessment movement, as a typical form of assessment that considers both the learning outcomes (the product)

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

and the developmental aspects (the process). A portfolio can be defined as a collection of works that demonstrate student competence (Paulston, Paulston and Meyer, 1979): for example, one may find in a student's portfolio the five best texts produced during a course that focuses on writing skills. Of course, the writing tasks should comply with the conditions of authentic assessment – complexity and contextualisation. More importantly, the works to be included in the portfolio are selected by the learner, under the guidance of the teacher who must provide the learner with criteria for the selection of the material. In addition, the portfolio must include some indication of the reasons why these works have been selected. The construction of the portfolio involves self-reflection about learning and the development of metacognitive skills (Arter and Spandel 1992). The learner plays an active role in the assessment. The portfolio is also a dynamic object that changes over time and may be examined by various persons.

Wiggins deliberately understates the psychometric problems that may arise with the implementation of methods like portfolio construction within the authentic assessment approach. Although the evolution of validity theory and the limitations of psychometric instruments involve a redefinition of our criteria for constructing good assessment tools (Moss, 1995), test designers still have to maintain the quality of the "interpretative argument" (Shepard, 1993; Kane, 1992), which is the chain of inferences that are drawn, starting from the observed performance to the conclusions about an underlying competence that may be activated in a variety of situations. More recently, Kane, Crook and Cohen (1999) distinguish three steps in the inference process:

- The scoring of the observed performance – in the context of authentic assessment, performances are often assessed in a holistic way to make sure that the interactions between the various components of the competence are taken into account and to replicate real-life judgments which are generally global.
- The generalisation to a universe of tasks – observed scores provide valuable and useful information as long as they are based on a representative sample of tasks. The problem with authentic assessment is that contextualised tasks often carry specific features that cause sampling errors and make generalisation uncertain. Standardisation may help to solve this problem but it makes the task less authentic and may affect the third type of inference.
- The extrapolation to a target domain – the basic assumption that allows this last type of inference is that the skills and knowledge called for by the task are those that are necessary to produce different performances in situations that pertain to the domain. Authentic tasks are interesting because they are similar to the real performances that are included in the target domain. For example, the production of a personal letter that is sent to a real person under normal conditions involves a combination of skills and knowledge: human relations, planning, vocabulary, grammar, coherence, handwriting and so on.

## 4 Authenticity in language testing

Debates in language testing are at the crossroads of current educational measurement concerns and language teaching and learning concerns. As mentioned before, the problem of authenticity had already been raised among researchers and practitioners in the

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

field of language teaching and learning (Hoekje and Linnell, 1994). Bachman (1990: Ch. 4) makes a distinction between two approaches of authenticity in language testing.

- The first approach is the "real life approach". Since it is concerned with the extent to which a test performance replicates a non-test performance, this approach favours direct tests in which the learner is observed in normal life: for example, during the execution of professional activities in the work place. Since direct testing is often impractical, testing and assessment procedures have been developed within this approach to create situations that look like situations the test-taker could be faced with in real life. The most important features of the real situation must be reproduced in the test setting. Many interview procedures have been designed to measure oral interaction within this framework. Role-playing and simulation are used and it is expected that the test taker will produce the same type of performance as in a real interaction situation. Real-life approach has often been criticised as naive. Most of the authors agree with the idea that trying to replicate a real-life situation in a testing situation is an oversimplified view of authenticity because the test-taker does not choose to engage in normal language use but is asked to display knowledge or skill (Stevenson, 1985; Spolksy, 1985).
- Bachman calls the second approach "interactional ability". Authenticity is obtained in two ways. First the test designer has to ensure that the test method is compatible with the target language-use situation. This requirement is less stringent than the real-life approach principle, but it must be met. For example, a cloze test which requires the test-taker to fill the gaps in a text cannot be considered as an authentic test if the purpose is to certify whether or not the test-taker can deal with tourism situations. On this matter, let us remember Harrison (1983: 78): "Who wants to stay in a hotel so damp that parts of the notices on the wall are obscured by patches of mould?". However, a cloze test may be used as a placement tool if the syllabus includes exercises where the student must find the missing information in a text. Secondly, and perhaps more importantly, the test designer has to ensure that the task actually impels the test-taker to use the knowledge and skill which are supposed to be used in the situation. Authenticity arises when both the test method aspects and the abilities that are required for the task are compatible with the characteristics of the target language-use domain.

Bachman and Palmer (1996) introduce the notion of test usefulness which is a function of the qualities that should be maximised. Authenticity is among these six qualities:

- *Reliability*: reliability can be defined in terms of consistency of test results across different versions of the same test, different raters and different equivalent tasks.
- *Construct validity*: the construct validity is related to the accuracy of inferences that can be made in relation to the theoretical definition of the competence and the purpose of the test.
- *Authenticity*: Bachman and Palmer (1996: 23) define authenticity as "the degree of correspondence of the characteristics of a given language task to the features of target language use task" and favour the interactional ability approach.

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

Authenticity is a desirable quality of a test because test interpretation depends on the demonstration that performance that is observed is not specific to the test itself. From this point of view, one may question the relevance of not including authenticity in construct validity.

- *Interactivity*: interactivity is not defined here in terms of communicative transactions between the test-taker and other persons or with a machine, but in terms of personal involvement in the task. Based on Stern's second type of authenticity one may argue that personal involvement in the test task is a prerequisite, or at least one aspect of authenticity.
- *Impact*: this quality is related to consequential validity (Messick, 1989) or systemic validity (Frederiksen and Collins, 1989). The form and the content of the tests that are used affect the teaching approaches and the learning strategies. Positive impact is an important asset of authentic assessment as described by Wiggins.
- *Practicality*: as we all know, a test that is too long or too cumbersome will never be used. This is a major constraint for direct testing using the real life approach because an adequate replication of a real-life situation is sometimes impossible for practical reasons. Practicality is certainly one of the qualities that can be improved with the use of computers.

## 5 Adaptive testing

Now that the multiple aspects of authenticity have been discussed, let us examine computerised testing and see how authentic it can really be. First, it is important to underline that research on the use of computers in language testing has been almost entirely devoted to the application of the principles of adaptive testing. In a recent article, Brown tries to summarise the present research on language computerised testing and singles out some future directions in the field. However, Brown (1997: 44) admits that he decided to "examine the dominant issue of computer-adaptive testing in the educational measurement literature in an attempt to forecast some of the directions future research on computers in language testing might take". The major part of his article deals with adaptive testing. Although I have been personally involved in the development of adaptive language tests (Laurier, 1993), I do not agree with the domination of adaptive testing in the research on the use of computers for language testing and assessment, and I believe that the capability of computers to administer tests that are tailored to the learner's proficiency level has probably been over-emphasised. Nonetheless, adaptive testing is certainly a major direction in the field. The fact that the TOEFL is now currently available in a computerised version that makes use of refined adaptive procedures demonstrates the importance of adaptive testing.

Given the importance of adaptive testing, I will briefly explain how it works to show that these procedures do not necessarily improve the authenticity of a test and may even limit the development of authentic tests. Adaptive testing is probably one of the most interesting applications of a psychometric theory that is called the item response theory. Items can be defined as isolated tasks that measure a common dimension. For example, a comprehension question on a paragraph could be an item in a bank of items that share this same format and that could measure "reading in short contexts" at different

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

difficulty levels. Another example is an item bank that would consist of multiple-choice questions on verbs. The most demanding operation in the development of an adaptive test is the experimental administration of these items in order to collect data to determine the psychometric properties of each item. Various indices may be calculated and items that fit in the bank will be, at least, assigned a difficulty value. During the computerised adaptive administration, the program will always select the item that best matches the estimation of the learner's level. As the number of items increases the accuracy of the estimate is improved.

Among the advantages of adaptive testing, the following should be mentioned:

- Accuracy of the ability estimates, because only the most informative items are used;
- Test security, because the tests are different from one student to the other and no test material circulates;
- Individualised administration so that timed group sessions are no longer necessary;
- Immediate feedback of the general score that can be reported right after the test;
- Shorter tests and, as a result, a more economical test and less demanding operation;
- Less frustrating tests from the learner's viewpoint as the items that are presented are never too difficult or too easy.

However, if we refer to Bachman's test qualities, we cannot find anything in adaptive tests that will improve assessment authenticity. In fact, these advantages are often obtained at the expense of authenticity. The major threat to authenticity is the so-called assumption of unidimensionality which means that all the items that are included in a bank must measure the same ability (Canale, 1986). Since an authentic assessment task is complex and contextualised, it is very unlikely to be unidimensional: on the one hand, an aggregate of abilities that may vary from one task to the other is required for the execution of the task and, on the other hand, the context brings a specific dimension to the task that is not common to the other items in the bank. The other limitation, which also concerns most computer tests, is the difficulty in handling responses that are not scored as either right or wrong.

## 6 Direct testing

As the use of computers becomes widespread, the possibilities of using them for direct testing within the real-life approach increase. Many language-oriented activities in our daily life, in the workplace, in academic settings, in our leisure time, for our social life, are now done in a computerised environment. Although nobody ever seriously considered reading Shakespeare's entire works or *À la recherche du temps perdu* on the screen, we must admit that reading is a language-oriented activity that is very often accomplished with a computer. In fact, many information processing tasks involve reading texts that have been found on the Internet and sometimes downloaded or reading e-mail or software installation procedures and on-line help. The material may be read on the screen or printed. As a result, a lot of people probably read more on a computer screen

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

than on paper. This change is not only quantitative but also qualitative as new types of reading activities arise. It is interesting to observe that in the international study on adult literacy conducted in 1994 under the supervision of Educational Testing Service, the questionnaire did not include any reference to the role of computers in reading. Updating this questionnaire in now being planned for the 2002 study; everybody agrees that new categories must be added to take into account new forms of reading in order to measure literacy adequately. In this context, computerised tests may be direct tests.

Let's use the example of a certification test that would be intended to verify how well a secretary can operate in the second language. The secretary will have to use a word processor in the target language and be able to understand the commands, error messages, help screens and so on. S/he will have to type a text that has been written by someone else, run the grammar and spell checks and make the appropriate changes. The secretary may have to make reservations, buy some products, retrieve information using the Internet in the target language. S/he may have to check e-mail and answer some of the messages. All these tasks can be tested directly on the computer. Of course, this type of test requires an observer to assess the performance.

One must ensure that computer familiarity does not interfere with language competence. If the tasks are not biased because of computer familiarity and if the tasks are carefully planned, there should not be any method effect that could reduce test validity. Moreover, a careful content analysis of the task in addition to some correlational analysis of the scores with other types of tests may reveal that these tasks put into play knowledge and skills that are common to other tasks that are not necessarily accomplished in a computerised environment. In that case, the computer would provide a context to assess complex skills that would be more difficult to assess in another way and that would be part of a relatively large target domain.

Like Brown, Van der Linden, one of the most famous researchers in educational measurement, discusses mostly adaptive testing in his recent article on computerised testing in education. However, he finally acknowledges the promise of multimedia testing and admits that "these media have the power to represent real-life situations convincingly in the test or to simulate processes in the real time" (Van der Linden, 1999: 149). Then he writes: "The potential use in educational testing has not been fully explored but their impact on the validity of tests may be enormous". Van der Linden refers here to the use of multimedia as a way to enhance the context. A richer context for the task also means more authentic assessment.

For example, a writing task may be introduced with a visit to a website as a prompt. Semi-direct oral proficiency interviews such as the SPEAK Test (Educational Testing Service, 1990), which consists of a sequence of recorded questions that replicates a face-to-face interview could certainly be improved with computers to enhance the context. Visuals could help to represent the interlocutor and the physical environment; it would then be possible for the test-taker to meet more than one interlocutor and move from place to place during the administration of the test. We could even imagine a test of oral interaction that would integrate in a scenario various activities in which the test-taker is asked to listen to a presentation, ask questions to the presenter, summarise it for someone else, discuss some ideas in relation to the theme of the presentation with a group of people and so on. These proposals are not entirely new because many computer

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

language programs have been created following this design. However, most of them do not include a real assessment component other than a raw score, a variation of the hit and miss ratio game score.

Ensuring measurement quality with the use of computers to enhance the context is not easy. Test development must go beyond the simple replication of the important features of the real context and adopt the interactional ability approach as proposed by Bachman. It is important to specify what knowledge and skills are actually necessary to accomplish the tasks successfully. We have to make sure that the context does not promote the use of test strategies that are specific to the computerised environment. Asking the student to accomplish various tasks in different contexts enlarges the samples of tasks and improves the quality of inferences particularly for generalisations to the universe of tasks. From a Utopian perspective, we may certainly envision the application of virtual reality techniques to create contexts for the communicative task. A few years ago, Teasdale (1996) presented a test intended for airline pilots in which the test-taker had to understand and reply to messages from the control tower. One can imagine how this test would have been improved if virtual reality techniques had been used to simulate a real flight from the cockpit. Of course, one of the problems in this type of environment is the analysis of the test-takers' production. This is a two-step process. First, the message must be decoded. Although there is no problem decoding messages entered on the keyboard, voice recognition techniques and even conversions of handwritten text are still too unreliable for assessment purposes. Second, the production has to be rated. We could certainly imagine a future program that would analyse the quality of the discourse. However, proofreading tools that are presently available are not very useful for judgments about the quality of learners' productions.

## 7 Assessment data processing

For the moment, we must be realistic. We cannot expect, in a the near future, the development of educational programs that would be able to replicate teachers' or examiners' judgment about a performance. Computer programs are very effective in handling selected answer items such as multiple choice or short answers. These items may be very useful in particular situations where authenticity in not a major issue and not even desirable, like in placement testing or some types of diagnostic tests. However, open-ended questions are unavoidable for the assessment or testing to be authentic. Although they are not always reliable, human judgments can hardly be replaced by computers when dealing with constructed answers to open-ended questions. There are very promising applications of artificial intelligence principles and advanced measurement techniques for student modelling, inference engine rules or automated diagnoses but domains that have been examined so far are much simpler than language competence (Giardina and Laurier, 1999).

If we do not insist on the principle that the inference process must be computerised, there are many other ways the computers may be used 'authentically'. As mentioned earlier, the most typical tool that has been proposed in the movement of authentic assessment is the portfolio. Although some computer programs are now available to create electronic portfolios, the power of computers has been underestimated. The portfolio, as a dynamic collection of documents requires the learner to reorganise his or

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

her portfolio, modify the documents and publish them in different ways. In addition, particularly in a language class, the portfolio should not be restricted to text documents but should also comprise recordings and even pictures. Computers can help the learner to modify, organise and present his or her personal portfolio. The learner's samples of performances are stored in a personal database that may include a variety of work that has been done during the language course such as short written and audio messages, exercises, poems, excerpts of chat sessions and even web pages. The learner must try to demonstrate the development of his or her language competence in a portfolio, in printed or digital form, that could be presented not only to the teacher, but also to peers, parents, school directors, employers and so on.

Finally, whereas in authentic language assessment, computers are not likely, in the near future, to undertake the inference process, we cannot disregard the contribution of computers in the development of tests. Psychometricians have been pioneers in the use of computers in education. As a matter of fact, the development of the Item Response Theory that underlies adaptive testing would not have been possible without the number-crunching capability of computers. Although the Item Response Theory focussed in the past on right or wrong answers, and more specifically on multiple choice items, interesting models for the analysis of graded responses that can be applied to the analysis of performances are now implemented in computer programs. A program called FACETS (Linacre, 1999) analyses different facets of a performance such as the task, the student level, the rater, the moment so that test design problems can be detected in the construction of authentic assessment tools. This program is now currently used to develop language performance tests for making important decisions (McNamara, 1996). Multidimensional analysis procedures have been improved and should be affordable shortly for the analysis of authentic assessment tasks so that we will be able to account for the multiple dimensions of language competence. All these complex psychometric analyses require a computer. In this context, even though the computer is not involved in either the administration or the scoring of the test, its contribution at the development stage of authentic tasks is essential.

## 8 Conclusion

To conclude, I would like to pay tribute to specialists in applied linguistics whose contribution has been important in the reflection on educational testing and assessment in general. Given the importance of assessment in our educational systems, the debate on authenticity forced educational measurement specialists to revisit their views. I hope that these new ideas will eventually lead to better language testing and assessment practices and more authentic tests. I think that computerised testing should go beyond adaptive testing and that computers will be useful in the design of more authentic language tests:

1. For direct testing of language-oriented tasks that are accomplished with a computer;
2. To enhance the context of the language tasks with the use of multimedia;
3. To be used by student in the construction of a portfolio;
4. To facilitate the psychometric analysis of authentic tasks.

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

## References

Arter, J. A. and Spandel, V. (1992) Using portfolios of students work in instruction and assessment (ITEMS modules), *Educational Measurement: Issues and Practices*, **12**(1): 36–44.

Bachman, L. F. (1990) *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. and Palmer, A. S. (1996) *Language Testing in Practice*. Oxford: Oxford University Press.

Brown, J. D. (1997) Computers in language testing: present research and some future directions. *Language Learning and Technology*, **1**(1): 44–59. http://polyglot.cal.msu/llt

Canale, M. (1986) The promise and threat of computerised adaptive assessment of reading comprehension. In: Stansfield, C. W. (ed.), *Technology and Language Testing*. Washington, DC: TESOL, pp. 29–46.

Educational Testing Service (1990) *Test of spoken English: manual for score users*. Princeton, NJ.

Frederiksen, J. R. and Collins, A. (1989) A system approach to educational testing. *Educational Researcher*, **18**(9): 27–32.

Giardina, M. and Laurier, M. (1999) Modélisation de l'apprenant et interactivité. *Revue des sciences de l'éducation*, **24**(3): 35–60.

Harrison, D. A. (1983) Communicative testing: jam tomorrow? In: Hughes, A. and Porter, D. (eds.), *Current Developments in Language Testing*. London: Academic Press, pp. 77–85.

Hoekje, B. and Linnell, K. (1994) Authenticity in language testing: evaluating spoken language test for international teaching assistants, *TESOL Quarterly*, **28**(1): 103–126.

Kane, M. T. (1992) An argument-based approach to validity. *Applied Psychological Measurement*, **6**(2): 125–160.

Kane, M., Crooks, T. and Cohen, A. (1999) Validating Measures of Performance. *Educational Measurement: Issues and practices*, **18**(2): 5–17.

Krashen, S. D. (1982) *Principle and Practice in Second Language Acquisition*. Oxford: Pergamon.

Laurier, M. (1993) *L'informatisation d'un test de classement en langue seconde*. Québec: International Center for Research on Language Planning.

Linacre, J. M. (1999) Computerised Educational Testing. In: Masters, G. M. and Keeves, J. P. (eds.), *Advances in Measurement in Educational Research and Assessment*. Oxford, Amsterdam: Pergamon, pp. 244–253.

McNamarra, T. (1996) *Measuring Second Language Performance*. London, New York: Longman.

Messick, S. (1989) Validity. In: Linn, R. L. (ed.), *Educational Measurement, 3rd ed.* New York: Macmillan/ACE, pp. 13–104.

Messick, S. (1994) The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher*, **23**(2): 13–23.

Morrow, K. (1979) Communicative language testing: revolution or evolution?. In: Brumfit, C. J. and Johnson, K. (eds.), *The Communicative Approach to Language Teaching*. Oxford: Oxford University Press, pp. 143–157.

Moss, P. A. (1995) Themes and variations in validity theory. *Educational Measurement: Issues and Practices*, **14**(2): 5–12.

Paulston, F. L., Paulston, P. P. and Meyer, C. A. (1979) What makes a portfolio a portfolio? *Educational Leadership*, **48**(5): 60–63.

Shepard, L. A. (1993) Evaluating test validity. In: Darling-Hammond, L. (ed.), *Review of Research in Education*, **19**. Washington, DC: AERA.

Spolsky, B. (1985) The limits of authenticity in language testing. *Language Testing*, **2**(1): 21–40.

Stern, H. H. (1982) Le concept de l'authenticité dans l'enseignement des langues vivantes : tentative de synthèse. *Bulletin of CALL*, **4**(2): 83–97.

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press

Stevenson, D. K (1985) Authenticity, validity and a tea party. *Language Testing*, **2**(1): 41–47.
Teasdale, A. (1996) Content validity in tests for well defined LSP domains: an approach to defining what is being tested. In: Milanovic, M. and Saville, N. (eds.), *Performance Testing, Cognition and Assessment*. Cambridge: Cambridge University Press, pp. 44–67.
Terwillinger, J. (1997) Semantics, psychometrics, and assessment reform: a close look at "authentic assessment". *Educational Researcher*, **26**(8): 24–27.
Van der Linden, W. J. (1999) Computerised educational testing. In: Masters, G. M. and Keeves, J. P. (eds.), *Advances in Measurement in Educational Research and Assessment.* Oxford, Amsterdam: Pergamon, pp. 138–150.
Wesche, M. B. (1992) Performance testing for work-related second language assessment. In: Shohamy, E. and Walton, R. (eds.), *Language Assessment for Feedback: Testing and other strategies*. Washington, DC: National Foreign Language Center, pp. 103–122.
Widdowson, H. G. (1979) *Exploration in Applied Linguistics*. Oxford: Oxford University Press.
Wiggins, G. (1989) A true test: toward more authentic and equitable assessment. *Phi Delta Kappa*, **20**(3): 703–713.
Wiggins, G. (1998) An exchange of views on Semantics, psychometrics, and assessment reform: a close look at 'authentic assessment', *Educational Researcher,* **27**(6): 19–20.
Wiggins, G. (1993) *Assessing Student Performance.* San Francisco: Jossey-Bass.

**Michel Laurier** is a professor at the University of Montreal and a specialist in second language testing and assessment. He graduated from the Ontario Institute of Studies in Education where he began his research on adaptive testing. He developed several language tests. He is currently the Director of the Heritage Language Center in Montreal.

https://doi.org/10.1017/S0958344000001014 Published online by Cambridge University Press