Authenticity in language testing: some outstanding questions

Jo A. Lewkowicz *English Centre, University of Hong Kong*

This article is divided into two main sections. Following the introduction, Section II takes a look at the concept of authenticity and the way this notion has evolved in language testing and more recently in general education. It argues that, although our understanding of the notion of authenticity has developed considerably since it was first introduced into the language testing literature in the 1970s, many questions remain unanswered. In an attempt to address one of the outstanding issues, Section III presents a study looking at the importance of authenticity for test takers. It shows that test takers are willing and able to identify the attributes of a test likely to affect their performance. However, these attributes do not necessarily include authenticity which has hitherto been considered an important test attribute for all stakeholders in the testing process. The article concludes that much more research is needed if the nature and role of authenticity in language testing is to be fully understood.

I Introduction

Any attempt at the characterization of authenticity in relation to assessment theory or practice needs to acknowledge that the notion of authenticity has been much debated both within the fields of applied linguistics as well as general education. In applied linguistics the notion emerged in the late 1970s at the time when communicative methodology was gaining momentum and there was a growing interest in teaching and testing ‘real-life’ language. In general education, on the other hand, it took more than another decade before the notion gained recognition. Since then there has been much overlap in the way the term has been perceived in both fields, yet the ‘debates’ have remained largely independent of each other even to the extent that, in a recent article, Cumming and Maxwell (1999: 178) attribute ‘[t]he first formal use of the term ‘authentic’ in the context of *learning and assessment* . . . to Archbald and Newmann (1988)’ (my emphasis).

Although many different interpretations of authenticity and authentic assessment have emerged, one feature of authenticity upon

Address for correspondence: Jo Lewkowicz, Associate Professor, English Centre, The University of Hong Kong, Pokfulam Road, Hong Kong; e-mail: jolewkow@hkusua.hku.hk

Language Testing 2000 17 (1) 43–64 0265-5322(00)LT171OA © 2000 Arnold

which there has been general agreement over time is that it is an important quality for test development which 'carries a positive charge' (Lynch, 1982: 11). Morrow (1991: 112), in his discussions of communicative language testing, pointed to 'the overriding importance of authenticity', while for Wood (1993) it is one of the most important issues in language testing. Wood (1993: 233) has proposed that there are two major issues – those of validity and reliability – and that they 'coalesce into one even greater issue: authenticity vs. inauthenticity'. Bachman and Palmer (1996), too, see authenticity as crucial. They argue that it is 'a critical quality of language tests' (p. 23), one that 'most language test developers implicitly consider in designing language tests' (p. 24). Authenticity is also pivotal to Douglas' (1997) consideration of specific purpose tests in that it is one of two features which distinguishes such tests from more general purpose tests of language (the other feature being the interaction between language knowledge and specific purpose content knowledge). The same positive sentiment is echoed in the field of general education where authentic assessment has been 'embraced enthusiastically by policy-makers, curriculum developers and practitioners alike', being seen as 'a desirable characteristic of education' (Cumming and Maxwell, 1999: 178).

Despite the importance accorded to authenticity, there has been a marked absence of research to demonstrate this characteristic. It is clear that authenticity is important for assessment theorists, but this may not be the case for all stakeholders in the testing process. It is not known, for example, how test takers perceive authenticity. It may be that authenticity is variably defined by the different stakeholders. It is also unclear whether the presence or absence of authenticity will affect test takers' performance. Bachman and Palmer (1996: 24) suggest that authenticity has a potential effect on test takers' performance. However, this effect is among those features of authenticity which have to be demonstrated if we are to move from speculation about the nature of authenticity to a comprehensive characterization of the notion. Before this can be achieved, a research agenda informed by our current understanding of authenticity and an identification of the unresolved issues needs to be drawn up. To this end, this article first reviews the current authenticity debate within the field of language testing and relates it to the debate in general education. It identifies a range of questions which need to be attended to for a better understanding of authenticity to be achieved. The article then goes on to outline in some detail a study which sets out to address one of the questions identified, and to suggest that there is not only a need for, but also value in, a systematic investigation of authenticity.

II The authenticity debate

1 The early debate

In applied linguistics the term ‘authenticity’ originated in the mid 1960s with a concern among materials writers such as Close (1965) and Broughton (1965) that language learners were being exposed to texts which were not representative of the target language they were learning. Close (1965), for example, stressed the authenticity of his materials in the title of his book *The English we use for science*, which utilized a selection of published texts on science from a variety of sources and across a range of topics. Authenticity at the time was seen as a simple notion distinguishing texts extracted from ‘real-life’ sources from those written for pedagogical purposes.

It was not until the late 1970s that Widdowson initiated a debate on the nature of authenticity. He introduced the distinction between ‘genuineness’ and ‘authenticity’ of language arguing that:

> Genuineness is a characteristic of the passage itself and is an absolute quality. Authenticity is a characteristic of the relationship between the passage and the reader and has to do with appropriate response. (Widdowson, 1978: 80)

Widdowson (1979: 165) saw genuineness as a quality of all texts, while authenticity as an attribute ‘bestowed’ on texts by a given audience. In his view, authenticity was a quality of the outcome present if the audience could realize the author’s intentions which would only be possible where the audience was aware of the conventions employed by the writer or speaker (Widdowson, 1990). He argued that genuine texts would only be considered authentic after undergoing a process of authentication, a process which he suggested may only be truly accessible to the native speaker. He failed to account for the way language learners could progress towards being able to authenticate texts, or to describe the native speaker. However, in distinguishing between genuineness and authenticity, Widdowson drew attention to the importance of the interaction between the audience and the text and hence to the nature of the outcome arising from textual input.

The distinction between genuine and authentic language was not readily accepted (a point recently lamented by Widdowson himself; Widdowson, 1998), and the discussion of authenticity remained for some time focused on the nature of authentic input. This was equally true in language teaching as it was in the field of language testing. Those advocating change to pre-communicative testing practices (such as Rea, 1978; Morrow 1978; 1979; 1983; 1991; Carroll, 1980) equated authenticity with what Widdowson identified as genuine input and focused on the need to use texts that had not been simplified and tasks that simulated those that test takers would be expected to perform in ‘the real world’ outside the language classroom.

This understanding of authenticity, detailed in Morrow’s groundbreaking report of 1978, gradually began to filter through to language testing practices of the 1980s and 1990s. In 1981, for example, in response to Morrow’s (1978) report, the Royal Society of Arts introduced the *Communicative Use of English as a Foreign Language* examination. This was the first large-scale test to focus on students’ ability to use language in context (language use) rather than their knowledge about a language (language use) (Hargreaves, 1987); it was also the precursor to the *Certificates in Communicative Skills in English* introduced in the 1990s by the University of Cambridge Local Examination Syndicate (UCLES). Both these tests were premised on the belief that authentic stimulus material was a necessary component of any test of communicative language ability. The same premise informed the development of many other tests, particularly in situations where oral language was being assessed and simulations of real-life tasks became a part of direct tests of spoken ability (e.g. Oral Proficiency Interviews) and where language for specific purposes was being assessed, such as in the British Council/UCLES *English Language Testing Service* (ELTS) test battery (for more detail, see Alderson *et al.*, 1987).

This conceptualization of authenticity gave rise, however, to a number of theoretical and practical concerns. First, by equating authenticity with texts that had not been altered or simplified in any way, a dichotomy was created between ‘authentic texts’ that were seen as intrinsically ‘good’ and ‘inauthentic texts’ produced for pedagogic purposes which were seen as ‘inferior’. This dichotomy proved unhelpful since it tended to ignore a number of salient features of real-life discourse. Texts produced in the real world differ (inter alia) in complexity depending on their intended audience and the amount of shared information between the parties involved in the discourse. Not all native speakers necessarily understand all texts (Seliger, 1985). Learning to deal with simple texts may, therefore, be a natural stage in the learning process and one that students need to go through (Widdowson, 1979; Davies, 1984). Using such texts in a test situation may similarly be considered the most appropriate for the language level of the test takers and, hence, may be totally justified. In addition, every text is produced in a specific context and the very act of extracting a text from its original source, even if it is left in its entirety, could be said to ‘disauthenticate’ it since authenticity, according to Widdowson (1994: 386), is ‘non-transferable’. In a test situation where, as Davies (1988) points out, it may not be possible or even practical to use unadapted texts, an obvious dilemma arises.

How should such a text be regarded: authentic because it has been taken from the real world, or inauthentic as it has been extracted from its original context for test use?

Another area of concern related to the view that authentic test tasks were those which mirrored real-life tasks. Such tasks are, by their very nature, simulations which cannot give rise to genuine interaction. They can, at best, be made to look like real-life tasks (Spolsky, 1985). Test takers need to cooperate and be willing to abide by the 'rules of the game' if simulations are to be successful in testing situations, otherwise the validity and fairness of the assessment procedures remain suspect (Spolsky, 1985). In addition, real-life holistic tasks do not necessarily lend themselves to test situations. Only a limited number of such performance-type tasks can be selected for any given test; additionally, the question of task selection for generalizations to be made from test to non-test performance seems never to have been adequately resolved. Morrow (1979) suggested characterizing each communicative task by the enabling skills needed to complete it and then determining the tasks by deciding on which enabling skills should be tested. This approach, as Alderson noted (1981), assumed that enabling skills can be identified. It also encouraged the breaking down of holistic tasks into more discrete skills, which Morrow (1979) himself recognized as problematic since a 'candidate may prove quite capable of handling individual enabling skills, and yet prove quite incapable of mobilizing them in a use situation' (p. 153).

Throughout the 1980s the authenticity debate remained firmly focused on the nature of test input with scant regard being paid to the role test takers play in processing such input. The debate centred on the desired qualities of those aspects of language tests which test setters control, with advocates of authenticity promulgating the use of texts and tasks taken from real-life situations (Morrow, 1979; Carroll, 1980; Doye, 1991), and the sceptics drawing attention to the limitations of using such input and to the drawbacks associated with equating such input with real-life language use (Alderson, 1981; Davies, 1984; Spolsky, 1985).

2 A reconceptualization of authenticity

In language teaching the debate was taken forwards by Breen (1985) who suggested that authenticity may not be a single unitary notion, but one relating to texts (as well as to learners' interpretation of those texts), to tasks and to social situations of the language classroom. Breen drew attention to the fact that the aim of language learning is to be able to interpret the meaning of texts, and that any text which moves towards achieving that goal could have a role in teaching. He proposed that the notion of authenticity was a fairly complex one and that it was oversimplistic to dichotomize authentic and inauthentic materials, particularly since authenticity was, in his opinion, a relative rather than an absolute quality.

Bachman, in the early 1990s, appears to have built on the ideas put forward by Widdowson and Breen. He suggested that there was a need to distinguish between two types of authenticity: situational authenticity – that is, the perceived match between the characteristics of test tasks to target language use (TLU) tasks – and interactional authenticity – that is, the interaction between the test taker and the test task (Bachman, 1991). In so doing, he acknowledged that authenticity involved more than matching test tasks to TLU tasks: he saw authenticity also as a quality arising from the test takers' involvement in test tasks. Bachman (1991) appeared, at least in part, to be reaffirming Widdowson's notion of authenticity as a quality of outcome arising from the processing of input, but at the same time pointing to a need to account for 'language use' which Widdowson's unitary definition of genuineness did not permit.

Like Breen (1985), Bachman (1990; 1991) also recognized the complexities of authenticity, arguing that neither situational nor interactional authenticity was absolute. A test task could be situationally highly authentic, but interactionally low on authenticity, or vice versa. This reconceptualization of authenticity into a complex notion pertaining to test input as well as the nature and quality of test outcome was not dissimilar to the view of authenticity emerging in the field of general education. In the United States, in particular, the late 1980s / early 1990s saw a movement away from standardized multiple-choice tests to more performance-based assessment characterized by assessment tasks which were holistic, which provided an intellectual challenge, which were interesting for the students and which were tasks from which students could learn (Carlson, 1991: 6). Of concern was not only the nature of the task, but the outcome arising from it. Although there was no single view of what constituted authentic assessment, there appears to have been general agreement that a number of factors would contribute to the authenticity of any given task. (For an overview of how learning theories determined interpretation of authentic assessment, see Cumming and Maxwell, 1999.) Furthermore, there was a recognition, at least by some (for example, Anderson *et al.* (1996), cited by Cumming and Maxwell, 1999), that tasks would not necessarily be either authentic or inauthentic but would lie on a continuum which would be determined by the extent to which the assessment task related to the context in which it would be normally performed in real-life. This construction of authenticity as being situated within a specific context can be compared to situational authenticity discussed above.

3 A step forward?

The next stage in the authenticity debate appears to have moved in a somewhat different direction. In language education, Bachman in his work with Palmer (1996) separated the notion of authenticity from that of interactiveness, defining authenticity as ‘The degree of correspondence of the characteristics of a given language test task to the features of a TLU task’ (Bachman and Palmer, 1996: 23). This definition corresponds to that of situational authenticity, while interactiveness replaced what was previously termed interactional authenticity. The premise behind this change was a recognition that all real-life tasks are by definition situationally authentic, so authenticity can only be an attribute of other tasks, that is, those used for testing or teaching. At the same time, not all genuine language tasks are equally interactive; some give rise to very little language. However, authenticity is in part dependent on the correspondence between the interaction arising from test and TLU tasks; regarding the two as separate entities may, therefore, be misleading. Certainly, Douglas (2000) continues to see the two as aspects of authenticity, arguing that both need to be present in language tests for specific purposes.

To approximate the degree of correspondence between test and TLU tasks – that is, to determine the authenticity of test tasks – Bachman and Palmer (1996) proposed a framework of task characteristics. This framework provides a systematic way of matching tasks in terms of their setting, the test rubrics, test input, the outcome the tasks are expected to give rise to, and the relationship between input and response (for the complete framework, see Bachman and Palmer, 1996: 49–50). The framework is important since it provides a useful checklist of task characteristics, one which allows for a degree of agreement among test developers interested in ascertaining the authenticity of test tasks. It takes into account both the input provided in a test as well as the expected outcome arising from the input by characterizing not only test tasks but also test takers’ interactions with these.

4 Outstanding questions

Operationalizing the Bachman and Palmer (1996) framework does, however, pose a number of challenges. To determine the degree of correspondence between test tasks and TLU tasks, it is necessary to ‘first identify the critical features that define tasks in the TLU domain’

(Bachman and Palmer, 1996: 24). How this is to be achieved is not clear. Identifying critical features of TLU tasks appears to require judgements which may be similar to those needed to identify enabling skills of test and non-test tasks. Once such judgements have been made, test specifications need to be implemented and, in the process of so doing, the specifications may undergo adjustment. This is particularly likely to happen during test moderation when, as recent research has revealed (Lewkowicz, 1997), considerations other than maintaining a desired degree of correspondence between test and non-test task tend to prevail. It must be remembered that test development is an evolutionary process during which changes and modifications are likely to be continually introduced. Such changes may, ultimately, even if unintentionally, affect the degree of correspondence between the test tasks and TLU tasks. In other words, the degree of authenticity of the resultant test tasks may fail to match the desired level of authenticity identified at the test specification stage.

Whether in reality such differences in the degree of correspondence between a test task and TLU tasks are significant remains to be investigated. It is possible that if one were to consider all the characteristics for each test task in relation to possible TLU tasks (a time-consuming process), then the differences in authenticity across test tasks might be negligible. Some tasks could display a considerable degree of authenticity in terms of input while others could display the same degree of authenticity only in terms of output, situation or any combination of such factors. None would feature as highly authentic in terms of rubric since this is likely to 'be a characteristic for which there is relatively little correspondence between language use tasks and test tasks' (Bachman and Palmer, 1996: 50).

The above issues, all of which relate to the problem of identifying critical task characteristics, give rise to a number of unresolved questions:

1) Which characteristics are critical for distinguishing authentic from non-authentic test tasks?
2) Are some of these characteristics more critical than others?
3) What degree of correspondence is needed for test tasks and TLU tasks to be perceived as authentic?
4) How can test developers ensure that the critical characteristics identified at the test specification stage are present in the resultant test tasks and not 'eroded' in the process of test development?

An underlying assumption which underpins the Bachman and Palmer framework is that TLU tasks can be characterized. This, however, may not always be possible or practical. In situations where learners have homogeneous needs and where they are learning a language for specific purposes, identifying and characterizing the TLU domain may be a realistic endeavour. Douglas (2000) suggests this to be the case. However, in circumstances where learners' needs are diverse and test setters have a very large number of TLU tasks to draw upon, such characterization of all TLU tasks may be unrealistic. Even if such a characterization were possible, it may not necessarily prove useful. The large number of TLU tasks characterized could ensure a level of authenticity for most test tasks selected, since the larger the number of TLU tasks to choose from, the more likely it is that there would be a level of correspondence between the test tasks and the TLU domain. This leads to the following questions:

5) Can critical characteristics be identified for all tests, that is, general purpose as well as specific purpose language tests?
6) If so, do they need to be identified for both general and specific purpose tests?

A third set of questions relates to test outcome: whether test tasks that correspond highly to TLU tasks in terms of task characteristics are perceived as authentic by stakeholders other than the test developers. There has been some research in this area to suggest that end-users may prove useful informants for determining the degree to which test tasks are perceived as authentic. In a study investigating oral discourse produced in response to prompts given as part of the Occupational English Test for Health Professionals, Lumley and Brown (1998) found that their professional informants perceived the tasks set as authentic. However, they also found that the tasks gave rise to a number of problems which restricted the authenticity of the language produced, that is, of the test outcome. They found that the role cards given to the test takers provided insufficient background information about 'their patient'. As a result, when discussing the patient's condition with the examiner (playing the role of a concerned relative), the test takers failed to sound convincing and authoritative. This would suggest that authenticity is made up of constituent parts such as authenticity of input, purpose and outcome, leading to the questions:

7) What are the constituents of test authenticity, and are each of the constituents equally important?
8) Does the interaction arising from test tasks give rise to that intended by the test developers?
9) To what extent can/do test tasks give rise to authentic-sounding output which allow for generalizations to be made about test takers' performance in the real world?

The final set of questions to be considered relate to stakeholder perceptions of the importance of authenticity. It has already been suggested that the significance of authenticity may be variably perceived among and between different groups of stakeholders. It is, for example, possible that perceived authenticity plays an important role in test takers' performance, as Bachman and Palmer (1996) propose. However, it is conceivable that authenticity is important for some – not all – test takers and only under certain circumstances. It is equally possible that authenticity is not important for test takers, but it is important for other stakeholders such as teachers preparing candidates for a test (see Section III). We need to address the following questions if we are to ascertain the importance of test authenticity:

10) How important is authenticity for the various stakeholders of a test?
11) How do perceptions of authenticity differ among and between different stakeholders of a test?
12) Does a perception of authenticity affect test takers' performance and, if so, in what ways?
13) Does the importance attributed to authenticity depend on factors such as test takers' age, language proficiency, educational level, strategic competence or purpose for taking a test (whether it is a high or low stakes test)?
14) Will perceived authenticity impact on classroom practices and if so, in what way(s)?

In relation to the final question (14), it is worth noting the marked absence of authenticity in discussions of washback (the impact of tests on teaching). The close tie drawn between authentic achievement and authentic assessment in educational literature implies a mutual dependence. Cumming and Maxwell (1999) go as far as to suggest that there is a tension between four factors – learning goals, learning processes, teaching activities and assessment procedures – all of which are in 'dynamic tension' and 'adjustment of one component requires sympathetic adjustment of the other three' (p. 179). Yet, literature on washback in applied linguistics fails to acknowledge this relationship. Wall (1997), for example, in her overview of washback does not mention the potential of test authenticity on classroom practices. Similarly, Alderson *et al*. (1995) – in considering the principles which underlie actual test construction for major examination boards in Britain – do not identify authenticity as an issue.

Authenticity, as the above overview suggests, has been much debated in the literature. In fact, there have been two parallel debates on authenticity which have remained largely ignorant of each other.

Discussions within the field of applied linguistics and general education – as Lewkowicz (1997) suggests – need to come closer together. Furthermore, such discussions need to be empirically based to inform what has until now been a predominantly theoretical debate. The questions identified earlier demonstrate that there is still much that is unknown about authenticity. As Peacock (1997: 44) has argued with reference to *language teaching*: ‘research to date on this topic is inadequate, and … further research is justified by the importance accorded authentic materials in the literature’. This need is equally true for *language testing*, which is the primary focus of this article.

One aspect of authenticity which has been subject to considerable speculation, but which has remained under-researched, is related to test takers’ perceptions of authenticity. The following study was set up to understand more fully the importance test takers accord to this test characteristic, and to determine whether their perceptions of authenticity affect their performance on a test.

III The study

1 The subjects

A group of 72 first-year students from the University of Hong Kong were identified for this study. They were all first-year undergraduate students taking an English enhancement course as part of their degree curriculum. The students were Cantonese speakers between 18 and 20 years of age. All had been learning English for 13 or more years.

2 The tests

The students were given two language tests within a period of three weeks. The two tests were selected because they were seen as very different in terms of their authenticity. The test administered first to all the students was a 90-item multiple-choice test based on a TOEFL practice test. It was made up of four sections: sentence structure (15 items), written expression (25 items), vocabulary (20 items) and reading comprehension (30 items). The students were familiar with this type of test as they had taken similar multiple-choice tests throughout their school career; additionally, part of the Use of English examination, which is required for university entrance, is made up of multiple-choice items.

The second test administered was an EAP (English for academic purposes) test which, in terms of Bachman and Palmer’s (1996) test task characteristics, was perceived as being reasonably authentic. Students took the test at the end of their English enhancement course, the course being designed to help them cope with their academic studies (English being the medium of instruction at the University). It was an integrated, performance test assessing many of the skills taught on the course. In terms of test type, students were somewhat familiar with integrated tests since one of the papers of the Use of English examination (the Work and Study Skills Paper) is integrated in nature. This EAP test was also made up of four sections:

1) listening and note-taking;
2) listening and writing;
3) reading and writing; and
4) synthesizing, selecting and organizing information from the earlier parts of the test to respond to a written prompt.

In the first part of the test, students listened to two extracts from academic-type lectures. The purpose of this task was to provide the students with information, and it was, therefore, not assessed. In the second part, students used their notes from the listening to write a summary of each of the extracts. The third part was based on two reading extracts of varying length. The first was a very short extract from a magazine article. The students were required to summarize the main point using academic tone and then comment on the extract's relevance to Hong Kong. The second extract was longer (approximately 450 words), taken from an academic journal article. Students were required to complete some notes on a specified aspect of the text. The final part of the test required students to write an academic essay using relevant information from the listening extracts and the readings. The aim of this final part was to have students integrate information from a variety of sources in a coherent manner and then comment on the subject in question.

3 *The procedure*

Immediately after each of the tests, while the students still had the test-taking experience in mind, they were asked to complete a questionnaire. Each questionnaire was designed to elicit the following information:

1) students' perceptions of what each section of the test was assessing;
2) students' perceptions of how they had performed on the test;
3) students' opinions of how well their performance reflected their ability to use English in an academic context;
4) students' perceptions of the advantages and disadvantages of the test type.

In addition, after the second test, students were asked to compare the two tests in terms of:

1) which in their view was a better indicator of their ability to use English in an academic context;
2) which they considered more accurately assessed what they had been taught in their enhancement classes.

The second questionnaire which included points (5) and (6) above, is given in Appendix 1.

One of the primary purposes of this questionnaire was to elicit whether students perceived authenticity as an important characteristic. However, direct, structured questions pertaining explicitly to test authenticity may not have been understood. They may also have suggested to students ideas which otherwise they may not have thought of. It was therefore decided to use an open-ended, unstructured question which avoided all use of jargon or complex terminology (question 4).

The other questions of interest here, for which results are reported in this article, are those which asked students to compare the two test types and their performance on these (questions 5 and 6). Both questions were open-ended asking students to explain their answers. The responses to all three of these questions were collated and the frequency of responses, where relevant, were compared using chi-square. Furthermore, the responses for students scoring in the top third for each test were compared to those scoring in the bottom third to ascertain whether performance on the test affected students' responses.

4 *Results of the study*

a Identification of test attributes: To determine which test features were important for students, question 4 of the questionnaire asked respondents to enumerate what they saw as the advantages and disadvantages of each of the tests. The first point to note is that the majority of the respondents (approximately 60%) appeared willing to identify at least one advantage and one disadvantage of each of the tests. This finding is confirmed by a statistical analysis of the frequency with which respondents identified advantages/ disadvantages for each test type; this shows that significantly more respondents identified than failed to identify test attributes (Table 1). A chi-square test of the differences in the frequencies with which students failed to identify advantages and disadvantages across the two test types shows no significant difference (Table 2). This finding further confirms that students were equally able to comment on the positive and negative attributes of both tests.

Table 1 Frequency with which respondents identified advantages and disadvantages of the two test types ($n = 72$)

Test type	Attribute	Number who identified	Chi-square	*df*	*p*
MC	Positive	65	46.722	1	.000
MC	Negative	46	5.556	1	.018
EAP	Positive	58	26.889	1	.000
EAP	Negative	50	10.889	1	.001

Notes: MC = Multiple-choice test; EAP = English for academic purposes test (an integrated performance test)

Table 2 Comparison of the frequency with which students failed to identify the advantages and disadvantages of the two test types

Attribute	Number who failed to identify (MC)	Number who failed to identify (EAP)	Chi-square	*df*	*p*
Advantage	7	14	2.732	1	.098 (n.s.)
Disadvantage	26	22	.500	1	.480 (n.s.)

Respondents noted a range of advantages and disadvantages of each of the tests. Among the positive attributes identified were the comprehensive nature of the EAP test (noted by 17% of respondents) and the apparent usefulness of the multiple-choice test (noted by 36% of respondents). Among the negative attributes were the amount of writing involved in the EAP test (noted by 25% of respondents) and the fact that the multiple-choice test was not interesting (noted by 8% of respondents). However, few respondents noted test features which could be identified as pertaining to authenticity. Of the 72 respondents, only 9 (12.5%) noted explicitly that an advantage of the EAP test was that it required responses similar to those expected of them in academic study. A further 10 respondents saw the writing component as an advantage, but failed to say why it was an advantage. This feature has, however, been loosely interpreted as pertaining to the authenticity of the test. Even so, the chi-square test – comparing the frequency with which the feature was noted with the frequency with which it was not noted – reveals that this difference was statistically significant at $p = .000$ (Table 3), suggesting authenticity was not an important feature for respondents. The frequency with which 'lack of authenticity' on the multiple-choice test was noted as a disadvantage was higher, with 26 respondents identifying this as a negative attribute of the test. However, when compared with the frequency with which this feature of the test failed to be noted, the difference

Table 3 Frequency with which authenticity/lack of authenticity was noted

Test type	Number who identified	Chi-square	*df*	*p*
EAP	19	5.556	1	.000
MC	26	16.056	1	.018

Note: see Table 1 for abbreviations

was again statistically significant, at a level of p = .018, suggesting that lack of authenticity was also perceived as unimportant for the majority of respondents (Table 3).

Attributes which respondents identified as likely to affect their performance included their familiarity with the task type, their motivation to do well (with a number of students being more motivated during the multiple-choice test which did not count towards their final grade, but which they saw as good revision practice) and the type of outcome required. A number seemed discouraged by the amount of writing required in the integrated EAP test, with 18 respondents noting this as a disadvantage of the test. This compares with 19 respondents who noted the writing as an advantage and 35 who did not identify this test feature.

b Relationship between academic performance and test type: Question 5 set out to determine which of the two tests the respondents considered to be a better indicator of their ability to use English in an academic context, that is, which they considered more authentic in terms of assessing TLU. Opinions seemed to be divided. As Table 4 shows, the percentage of respondents who identified the multiple-choice test was very similar to the percentage who identified the EAP test. Also, 14% of respondents recognized the need for both test types.

Table 4 Students' perception of the test type which better assessed their ability to use English in an academic setting

	MC	EAP	Both	No response	Irrelevant response
Percentage of total (*n* = 72)	36	38	14	7	6
EAP: percentage of top third	26	52	22	0	0
EAP: percentage of bottom third	48	22	9	13	9
MC: percentage of top third	38	50	8	4	0
MC: percentage of bottom third	38	42	8	4	8

Note: see Table 1 for abbreviations

Students' choice of test appears to have been related to their performance. Those who scored in the top-third for either test were more likely to select the integrated test as better reflecting their ability to perform in English in an academic setting than those who performed in the bottom third (see Table 4). Among the reasons provided for selecting the multiple-choice test were that it assessed a wider range of skills and that the test was taken under less pressure, with some respondents noting that this helped them to perform to the maximum of their ability. In contrast, those who selected the integrated EAP test noted that the skills being assessed included writing tasks which were more relevant to an academic context.

c Relationship between teaching and testing: Responses to question 6 indicate that a considerable number of students failed to perceive a connection between what was tested and what had been taught in their English enhancement course. As Table 5 shows, while 38% of the students indicated that the multiple-choice test better assessed the skills taught, 42% indicated the integrated test better assessed what had been taught. The reasons for selecting one or other test in this instance were very similar to those for selecting which test better assessed students' ability to use English in an academic context. Those in favour of the multiple-choice test identified its comprehensiveness, its objectivity in terms of marking, and the fact that it did not include a listening component which some noted was not explicitly taught in the enhancement classes. Those in favour of the integrated test, also noted its comprehensiveness as well as the fact that it targeted productive skills.

Table 5 Students' perception of the test type which better assessed what they had been taught in their English classes

	MC	EAP	Both	No response	Irrelevant response
Percentage of total ($n = 72$)	38	42	7	8	6
EAP: percentage of top third	43	43	9	4	0
EAP: percentage of bottom third	39	35	0	13	13
MC: percentage of top third	38	42	13	0	8
MC: percentage of bottom third	35	46	4	8	8

Note: see Table 1 for abbreviations

5 Discussion

Most of the respondents were willing and able to identify test attributes for both test types. They identified a wide range of attributes, but none was identified consistently by a majority of respondents. Furthermore, few saw aspects of authenticity or lack of it as important. It is possible that students did not think of noting aspects of test relevance when completing the open-ended question and, if prompted, they would have agreed that authenticity was important. Yet, students were able to identify other attributes, both positive and negative of the two test types. Since students were specifically asked to focus on those aspects of the tests they considered important and likely to affect their performance (something they are unlikely to do overtly under most test conditions), it would seem plausible to conclude that authenticity was not a priority for the majority of the respondents. It was only seen as important by some of the respondents, and authenticity was only one of a number of attributes that may have been identified. It is possible that this attribute is taken for granted by many test takers and noted only when, contrary to their expectations, it is absent from the test. Furthermore, authenticity may be viewed by some as a disadvantage rather than an advantage of a test. This appears to be true for the group of respondents who saw the amount of writing during the EAP test as a negative attribute.

The results further suggest that students' perceptions of what a test is testing and how that relates to both TLU and what is taught depends on their performance on the tests. Those performing well seemed better able to recognize the connection between the EAP test and the language which they were required to use in their studies. They were also better able to see the relationship between what was being tested in the EAP test and what was taught in their enhancement course. The latter result is somewhat surprising as all the students had gone through the same EAP course. They did not necessarily have the same teacher, but the variability of responses within each group suggests that this result was not dependent the teacher. A more likely explanation is that some of the respondents associate language assessment with proficiency-type tests, regardless of what has been taught. They may view multiple-choice tests as authentic tests of language in contrast to tests of authentic language.

IV Conclusion

This study is limited in that it relied on students' self-perceptions of test qualities, and the researcher had no opportunity to carry out follow-up interviews with the students. It does, however, show that valuable insights can be obtained about authenticity from test-taking informants and that their response to test tasks may be much subtler and more pragmatic than testers might prefer to believe. The results of the study also raises the issue of the importance accorded to authenticity in the literature. They show that test takers' perceptions of authenticity vary. For some it is an important test attribute likely to contribute to their performance; for others, the attribute is only noticed when absent from a test. Authenticity would appear not to be universally important for test takers. On one hand these results are in line with Bachman and Palmer's (1996) notion that stakeholders' perceptions of test authenticity differ not only across but also between groups of stakeholders; at the same time they suggest that there may be a mismatch between the importance accorded to authenticity by language testing experts and other stakeholders in the testing process. Authenticity may be of theoretical importance for language testers needing to ensure that they can generalize from test to non-test situations, but not so important for other stakeholders in the testing process.

Since the nature of the input in terms of the tests and the teaching leading up to the test were the same, factors other than the correspondence between test and TLU tasks must have affected students' perceptions of the test tasks. In the same way as van Lier (1996) argues that authenticity of materials used in teaching contexts constitute only one set of conditions for authenticity to be discernible in the language classroom, so it would appear that test input constitutes only one set of conditions for authenticity to be discernible in language tests. Other conditions suggested by the data relate to test takers' language ability. Students who scored in the top third on either test were more likely than those who scored within the bottom third to identify the integrated test as more closely assessing their ability to use the target language. This indicates a relationship between perception and performance. There are probably further conditions which affect test takers' perceptions and the way they process test input and interpret expected outcome. The role and importance of any of these factors has yet to be ascertained in the processing of test input. It would, however, appear that no single condition by itself will allow a test to be perceived as authentic by all test takers. A range of conditions interacting with the test input will affect test takers' perceptions and help determine whether test tasks are considered authentic or not. A high degree of correspondence between test and target-language use tasks may be a necessary but insufficient condition for authenticity to be discerned. We need to investigate these other conditions if we are to inch closer to understanding authenticity. We need to do so through continued empirical research. We also need to extend the research agenda to other aspects of authenticity so that in the long run the questions posited in the first section of this article are systematically addressed. In that way the debate on authenticity will be moved forward from one that has been largely theoretical to one that is based on research findings.

V References

Alderson, J.C. 1981: Reaction to Morrow paper. In Alderson, J.C. and Hughes, A., editors, *Issues in language testing: ELT Documents 111*. London: The British Council.

Alderson, J.C., Clapham, C. and **Wall, D.** 1995: *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alderson, J.C., Krahnke, K. and **Stanfield, C.** 1987: *Reviews of English language proficiency tests*. Washington, DC: Teachers of English to Speakers of Other Languages.

Bachman, L. 1990: *Fundamental considerations in language testing*. Oxford: Oxford University Press.

Bachman, L. 1991: What does language testing have to offer? *TESOL Quarterly* 25, 671–704.

Bachman, L and **Palmer, A.** 1996: *Language testing in practice*. Oxford: Oxford University Press.

Breen, M. 1985: Authenticity in the language classroom. *Applied Linguistics* 6, 60–70.

Broughton, G. 1965: *A technical reader for advanced students*. London: Macmillan.

Carlson, D. 1991: Changing the face of testing in California. California Curriculum News Report, 16/3.

Carroll, B.J. 1980: *Testing communicative performance*. London: Pergamon.

Close, R.A. 1965: *The English we use for science*. London: Longman.

Cumming, J. and **Maxwell, G.** 1999: Contextualising authentic assessment. *Assessment in Education* 6, 177–94.

Davies, A. 1984: Validating three tests of English language proficiency. *Language Testing* 1, 50–69.

Davies, A. 1988: Communicative language testing. In Hughes, A., editor, *Testing English for university study: ELT Documents 127*. Oxford: Modern English Publications.

Douglas, D. 1997: Language for specific purposes testing. In Clapham, C. and Carson, D., editors, *Encyclopedia of language in education. Volume 7: Language testing and assessment*. Dordrecht: Kluwer Academic, 111–20.

Douglas, D. 2000: *Assessing language for specific purposes*. Cambridge: Cambridge University Press.

Doye, P. 1991: Authenticity in foreign language testing. In Anivan, S., editor, *Current developments in language testing*. Singapore: SEAMEO Regional Language Centre, 103–10.

Hargreaves, P. 1987: Royal Society of Arts: Examination in the Communicative Use of English as a Foreign Language. In Alderson, J.C., Krahnke, K. and Stanfield, C. *Reviews of English language proficiency tests*. Washington, DC: Teachers of English to Speakers of Other Languages.

Lewkowicz, J. 1997: Investigating authenticity in language testing. Unpublished PhD dissertation, University of Lancaster.

Lumley, T. and **Brown, A.** 1998: Authenticity of discourse in a specific purpose test. In Li, E and James, G., editors, *Testing and evaluation in second language education*. Hong Kong: The Language Centre, The University of Science and Technology, 22–33.

Lynch, A.J. 1982: 'Authenticity' in language teaching: some implications for the design of listening materials. *British Journal of Language Teaching* 20: 9–16.

Morrow, K. 1978: *Techniques for evaluation for a notional syllabus*. London: Royal Society of Arts.

Morrow, K. 1979: Communicative language testing: revolution or evolution? In Brumfit, C.J. and Johnson, K., editors, *The communicative approach to language teaching*. Oxford: Oxford University Press, 143–57.

Morrow, K. 1983: The Royal Society of Arts Examinations in the Communicative Use of English as a Foreign Language. In Jordan, R., editor, *Case studies in ELT*. London: Collins ELT, 102–07.

Morrow, K. 1991: Evaluating communicative tests. In Anivan, S., editor, *Current developments in language testing*. Singapore: SEAMEO Regional Language Centre, 111–18.

Peacock, M. 1997: The effect of authentic materials on the motivation of EFL learners. *English Language Teaching Journal* 51, 144–56.

Rea, P. 1978: Assessing language as communication. *MALS Journal* 3.

Seliger, H.W. 1985: Testing authentic language: the problem of meaning. *Language Testing* 2, 1–15.

Spolsky, B. 1985: The limits of authenticity in language testing. *Language Testing* 2, 31–40.

van Lier, L. 1996: *Interaction in the language curriculum: awareness, autonomy and authenticity*. London: Longman.

Wall, D. 1997: Impact and washback in language testing. In Clapham, C. and Carson, D., editors, *Encyclopedia of language in education. Volume 7: Language testing and assessment*. Dordrecht: Kluwer Academic, 291–302.

Widdowson, H. 1978: *Teaching language as communication*. Oxford: Oxford University Press.

Widdowson, H. 1979: *Explorations in applied linguistics*. Oxford: Oxford University Press.

Widdowson, H. 1990: *Aspects of language teaching*. Oxford: Oxford University Press.

Widdowson, H. 1994: The ownership of English. *TESOL Quarterly* 28, 377–89.

Widdowson, H. 1998: Context, community and authentic language. Paper presented at TESOL Annual Convention, Seattle, 17–21 March 1998.

Wood, R. 1993: *Assessment and testing*. Cambridge: Cambridge University Press.

Appendix 1 Questionnaire

English Centre
The University of Hong Kong

Instructions

Please take some time to complete all the questions on this questionnaire. Questions 1–4 refer to the End-of-Course EAS test you have just completed. Questions 5 and 6 refer to both this test and the Practice Test you completed before the end of the course. You may refer to the question paper while completing this questionnaire.

1. What do you think each of the following sections were testing?

Section B: ....

....

Section C: ....

....

Section D: ....

....

2. How well do you think you have performed on each of the following sections of this test? Tick the appropriate boxes.

	Section B	Section C	Section D
above 80%			
71–80%			
61–70%			
51–60%			
41–50%			
below 40%			

3. Indicate below how well you think your overall score on this test will reflect your ability to use English in your studies?

Very well x------x------x------x------x------x------x Not at all

4. Having completed this test, what do you think are the good and bad points of this type of test? Please list as many points as you can.

   Good Points *Bad Points*

5. Which test, the Practice Test or this End-of-Course Test, do you think is a better indicator of your ability to use English in your studies? Why?

6. Which test, the Practice Test or this End-of-Course Test, do you think better assesses what you have learned during your EAS classes? Give reasons for your choice.

Thank you for your co-operation.