PERGAMON

System 28 (2000) 483–497

SYSTEM

www.elsevier.com/locate/system

The 'communicative' legacy in language testing

Glenn Fulcher *

English Language Institute, University of Surrey, Guildford, Surrey GU2 7XH, UK

Received 15 December 1999; accepted 11 April 2000

Abstract

This article looks at the phenomenon of 'communicative'' language testing as it emerged in the late 1970s and early 1980s as a reaction against tests constructed of multiple choice items and the perceived over-emphasis of reliability. Lado in particular became a target for communicative testers. It is argued that many of the concerns of the communicative movement had already been addressed outside the United Kingdom, and that Lado was done an injustice. Nevertheless, the jargon of the communicative testing movement, however imprecise it may have been, has impacted upon the ways in which language testers approach problems today. The legacy of the communicative movement is traced from its first formulation, through present conundrums, to tomorrow's research questions.

Keywords: Language testing; Communicative language testing

1. What is 'communicative' language testing?

Spolsky (1976) suggested that the history of language testing could be divided into three distinctive periods: the pre-scientific, the psychometric-structuralist, and the psycholinguistic-sociolinguistic. Morrow (1979, p. 144) translated these periods into the Garden of Eden, the Vale of Tears and the Promised Land. The Promised Land was the advent of 'communicative' language testing (as Morrow christened it) at the end of the 1970s and the early 1980s.

But what was 'communicative' language testing? The approach was primarily a rejection of the role that reliability and validity had come to play in language testing, mainly in the United States during the 1960s. The key targets to attack for the new

* Tel.: +44-1483-259910; fax: +44-1483-259507.
E-mail addresses: g.fulcher@surrey.ac.uk (G. Fulcher).

0346-251X/00/$ - see front matter
PII: S0346-251X(00)00033-6

'communicative' language testers were the multiple-choice item as embodied in the Test of English as a Foreign Language (Spolsky, 1995), and the work of Lado (1961).

Morrow (1979, pp. 146–147) characterised reliability (as defined by Lado) as the search for objectivity (in the use of multiple choice items), conveniently ignoring the distinction that Lado made between 'reliability' and 'scorability' (Lado, 1961, p. 31). Morrow further claimed that with the exception of face validity (and possibly predictive validity) the whole concept of validity is circular because it only exists in terms of criteria, all of which are relative and based upon questionable assumptions. It was the task of the communicative language tester to re-define reliability and validity. In order to do this, early communicative language testers latched onto Davies' (1978) argument that there was a "tension" between reliability and validity, and defined validity as the parallelism of real-world activities and test tasks. This meant that validity involved making the test truer to an activity that would take place in the real world. Yet, reliability could only be increased through using objective item types like multiple choice. As Underhill (1982, p. 18) argued: "there is no real-life situation in which we go around asking or answering multiple choice questions." (It should also be remembered that at this time validity was perceived to be a quality of the test, which is no longer the case.)

Thus, the more validity a test had, the less reliability it had, and vice versa. This was expressed most clearly in Underhill (1982, p. 17) when he wrote:

> If you believe that real language use only occurs in creative communication between two or more parties with genuine reasons for communication, then you may accept that the trade-off between reliability and validity is unavoidable.

Similarly, Morrow (1979, p. 151) had argued that, "Reliability, while clearly important, will be subordinate to face validity." That is, the key criterion in identifying a good test is that it looks like a good one, the input appears to be "authentic", and the task or item type mirrors an act of communication in the real world (Morrow, 1982, pp. 56–57). Carroll (1982, p. 1) went as far as to say that, "The communicative approach stands or falls by the degree of real life, or at least life-like, communication that is achieved..."

It was argued that communicative tests would involve performance (speaking), and the performance would be judged subjectively, qualitatively and impressionistically, by a sympathetic interlocutor/assessor. It is not insignificant that Carroll's (1980) book was entitled "Testing Communicative Performance", rather than "Testing Communicative Competence", picking up the view first expressed by Morrow (1979, p. 151) that 'communicative tests' will always involve performance.

The buzz words of early communicative language testing soon became:

1. real life tasks;
2. face validity;
3. authenticity; and
4. performance.

In the journey to the promised land, Morrow (1979, p. 156) prophesied that, "...there is some blood to be spilt yet." Underhill (1982, p. 18) preached: "As ye teach, so shall ye test." The communicative movement soon developed an antipathy to statistical analysis and language testing research, in which the "cult of the language testing expert" was deplored (Underhill, 1987, p. 1), and "common sense" portrayed as more important than "the statistical sausage machine" (ibid, p. 105).[1] Morrow (1991, p. 116) wrote of language testing research that:

> At its worst, it might tempt us to paraphrase Oscar Wilde's epigram about the English upper classes and their fondness for fox hunting. Wilde spoke of 'The English country gentleman galloping after a fox — the unspeakable in full pursuit of the uneatable'. Language testing researchers may not be unspeakable; but they may well be in search of the unmeasurable.

Perhaps popular communicative language testing of this type was neither revolution nor evolution. It was rebellion against a perceived lack of concern for the individual, the human, the "subject" taking the test; it wished to remove the use of so-called "arcane pseudo-scientific jargon", and rejoice in the common sense of the classroom teacher who can tell whether a test is good or not by looking at it; it wished to "re-humanise" assessment. Indeed, Morrow (1991, p. 117) referred to "ethical validity", and asked whether testing would make a difference to the quality of life of the test-taker. He perceived this particularly in terms of washback. However, he also recommended that test-takers should be "genuinely involved in the process" of assessment — although what this might mean in practice was not explored.

2. Was this all so new?

The new communicative language testers believed that the promised land offered something different and better from everything that had gone before (Harrison, 1983). But as Spolsky (1995) and Barnwell (1996) remind us, the work of those who have gone before is always in danger of being forgotten (or even misrepresented), which is precisely what happened in the early communicative language testing movement.

Indeed, many of the calls for change are reminiscent of debates that had taken place decades before. Speaking tests had been common in the United States since the late 1800s, and there is early evidence of innovative task types, such as asking a learner to interpret between two raters (Lundeberg, 1929). Nevertheless, it is true to say that throughout the 1920s, the heyday of language test development (Spolsky, 1995, pp. 33–51), and into the 1960s, the focus of language testing was on delivering large numbers of tests each year within a rapidly expanding education system in the

[1] Even though "Common sense might be embarrassed in defending its position against the awkward fact which has been adduced, that even in the best regulated examinations one examiner occasionally differs from another to the extent of 50 percent." (Edgeworth, 1890, p. 661)

United States. The answer to this problem was the 'new-type' test made up of multiple choice items. The multiple choice item was born out of the need to produce tests on an industrial scale, and their use was perpetuated through the development of new automatic marking machines that were designed especially to process multiple choice items (Fulcher, 2000). Facts not taken into account by the communicative critics of large-scale tests.

Communicative language testers also seem to have thought that their concerns about multiple choice items were new, but this was far from the case. Mercier (1933, p. 382) was just one of the few early writers to express concern that the item format elicited "passive" rather than "dynamic" knowledge, which might limit the generalisability of scores to language use. Further, language testers from the 1920s did realise that it was important to have individualised tests of speaking using questions and conversational prompts (Wood, 1927), but that such tests were not practical when delivering tests on an industrial scale (Hayden, 1920).

Apart from practicality there was one further reason why speaking tests were not developed for large-scale use at the time. This was the deep concern about reliability, where the 'subjective' judgement of an individual scorer dictated test outcome. Nevertheless, in the United States the College Boards test in English as a Foreign Language used throughout the 1920s and 1930s included a 15-min oral interview as part of the test battery (Barnwell, 1996, p. 70). However, the language sample was graded for pronunciation, as this was considered more reliable than any other criteria that might have been used.

With the Second World War, however, the ability to communicate in a second language became the explicit goal of teaching programmes. These teaching programmes depended more upon continual assessment, and new tests were not immediately developed to rate the personnel being trained in foreign languages. However, the war experience was undoubtedly the beginning of the modern testing of speaking and 'communicative' testing (Fulcher, 1998a). Kaulfers (1944, p. 137), for example, wrote that tests:

> ...should provide specific, recognizable evidence of the examinee's readiness to perform in a life-situation, where lack of ability to understand and speak extemporaneously might be a serious handicap to safety and comfort, or to the effective execution of military responsibilities.

In the British tradition too, there had always been speaking tests. However, these may have put test-takers at more of a disadvantage, because they were based on reading aloud and understanding extensive literary texts about which they had to express opinions (Spolsky, 1995, pp. 205–205). Further, there was little of the concern for test reliability that had emerged in the United States. Into the 1950s and 1960s whilst British test developers were arguing over whether multiple choice items should be allowed into their tests, and whether the literature requirement should be brought more up to date, testing agencies in the United States were developing the first true communicative performance tests that were to become the models for future test design and development around the world.

The Foreign Service Institute (FSI) speaking test of 1956 was the first test of speaking ability that required conversation with a trained rater (Sollenberger, 1978). The greater part of the test had been developed as early as 1952, and in 1958 subscales of accent, comprehension, fluency, grammar and vocabulary were added (although not used for initial ratings). So successful was the FSI oral proficiency interview that its use spread to the CIA, the Defense Language Institute and the Peace Corps. By 1968 a standard speaking test was produced for all these organisations in the form of the Interagency Language Roundtable speaking test (Lowe, 1983). In the 1970s this approach spread to schools and colleges in the United States (Liskin-Gasparro, 1984), and many innovative programmes, including rater training, were developed and disseminated (Adams and Frith, 1979). The rating scales, as they eventually applied to schools and colleges, were developed by the American Council on the Teaching of Foreign Languages, published in 1982 (draft) and 1986 (final). The format and number of bands were a direct result of research into scale sensitivity conducted two decades earlier (Carroll, 1961).

In addition to these developments a great deal of research into the reliability and validity of these tests was undertaken. In the British context, under the grip of the communicative school, such research was not generally conducted because of faith in face validity. Indeed, as late as 1990 in the first reference to a reliability study using University of Cambridge Local Examination Syndicate tests, Bachman and Davidson (1990, pp. 34–35) report that reliability coefficients were not available for speaking tests because they were never double scored. This echoed fears expressed earlier by researchers such as Hamp-Lyons (1987, p. 19). Despite the publication of the ALTE Code of Practice in 1994, committing European examination boards to the production of appropriate data, such basic statistical information has yet to be published (Chalhoub-Deville and Turner, 2000).

3. Reinstating Lado

The preceding brief review shows that the communicative language testing movement in the late 1970s and early 1980s was an unusual blip in the development of language testing. As a primarily British phenomenon, it has left its mark on the policy and practice of British EFL boards in the focus on face validity, 'authentic' looking tasks, and a dearth of research — if not an actual disdain of research — as demonstrated by Alderson and Buck (1993). This situation was predictable partly because the communicative testing movement was essentially a rebellion particularly against reliability, as *perceived* in the work of Lado (1961). Yet, this perception was one that did not, and does not, match the content of Lado's work. With hindsight, one must wonder if British critics of the late 1970s had actually read Lado at all. Lado (p. 239) wrote that: "The ability to speak a foreign language is without doubt the most prized language skill, and rightly so…". He went on, however, to explain that the testing of speaking was the least developed area of language testing, and that this "…is probably due in part at least to a lack of clear understanding of what constitutes speaking ability or oral production." The construct of "speaking"

was under-defined. Lado was deeply concerned about correctly classifying students into language ability levels, and was thus interested in reliability. He argued that because of the complexity of language production and the *non-language factors* involved, reliability was difficult to obtain. Validity and reliability were inextricably bound together. He was also keenly aware of other problems, as this quotation (Lado, p. 240) makes clear:

> Speaking ability is described as the ability to express oneself in life situations, or the ability to report acts or situations in precise words, or the ability to converse, or to express a sequence of ideas fluently.... This approach produces tests that must range over a variety of situations to achieve validity, and then there is no assurance that the language elements of speaking have been adequately sampled. Scoring is done by means of a rating scale describing the type of responses to be expected and the score to assign to each.

Lado therefore argued that it was better to test speaking through the "language elements" that were needed to communicate. He was aware of the problem of sampling situations in performance tests in such a way that task could be matched to abilities, so that claims about abilities would be valid and generalisable to other situations (within what would now be called a domain behaviour paradigm). However, a speaking test would still involve the learner speaking, as it is necessary in Lado's view to test the ability to produce language at a speed that is comparable to that of a native speaker (Lado's definition of "fluency"). Many of the activity types listed by Lado (1961, pp. 244–245) are not dissimilar to those of Underhill (1987) in sections labelled "picture series", "sustained speech", and "conversation". The sustained speech is essentially a role play conducted between the rater and the test-taker. The difference lies in Lado's awareness of the measurement requirements, the link to careful task and scale design, and construct definition.

With sections entitled "Testing the Integrated Skills" and "Beyond Language: How to test cross-cultural understanding," Davies (1978, p. 133) was surely correct when he wrote "there is more to Lado than analytical tests...".

The British communicative testing movement of the late 1970s and early 1980s was therefore lobbying for tests and task types that were already being developed outside the United Kingdom (see also Lowe, 1976), and simultaneously hindering the pursuit of any systematic research agenda in language testing that could address Lado's worries. Particularly in the field of testing speaking, little research has been done within British testing organisations to seek answers to many of the key questions relating to reliability, rating scale design and construction, test-taker characteristics, task variance, test-method facets, generalisability, and construct definition. Nevertheless, the fact that the jargon of British communicative language testing spread rapidly has impacted upon the direction of language testing research. In what follows, I will look at the criteria of a communicative test set out in 1979, and consider how these criteria have been developed in ways that provide new insights into language testing.

4. What is a 'communicative' test?

Morrow (1979) claimed that there were specific criteria that could be used to tell if a test is communicative. In an early commentary, Alderson (1981, p. 48) argued that communicative testers had failed to show: (1) how traditional language tests (undefined) fail to incorporate these criteria; and (2) how the criteria could be incorporated into communicative language tests. Nevertheless, these criteria (and the associated buzzwords) have left a legacy in language testing research because of the wide acceptance of 'the communicative approach'. This, in turn, has impacted upon language testers, who have developed new ways of looking at language assessment as a result.

4.1. Communicative tests involve performance

1. *Performance*: test-takers should actually have to produce language.
2. *Interaction-based*: there will be actual "face-to-face oral interaction which involves not only the modification of expression and content...but also an amalgam of receptive and productive skills" (Morrow, 1979, p. 149).
3. *Unpredictability*: Language use in real-time interaction is unpredictable.

Spolsky (1995) has shown that performance tests, in which learners take part in extended discourse of some form, have been in use for decades, if not centuries. The assumption underlying the performance tests advocated and developed in the 1980s was that the observation of behaviour that mirrored 'real-world communication' would lead to scores that would indicate whether the learner could perform in the real world. This 'real world' involves interaction, unpredictability[2], and integration of skills. The test and the criterion (real-world communication) are seen to be essentially the same, which led to the pre-eminent position of sampling and content analysis as the primary approach to test validation in English for Academic Purposes (EAP) testing (Fulcher, 1999). However, language testers are not usually interested in making a prediction from a test task only to the same task in the 'real world' — even though this is in itself an inference in need of validation. From a sample performance, the inference we need to make is usually to a range of potential performances, often as wide as "able to survive in an English medium University". For this, the performance is merely the vehicle for "getting to" underlying abilities, which are hypothesised to enable a range of behaviours relevant to the criterion situation(s) to which the tester wishes to predict. The requirement that we validate inferences drawn from scores, which are abstractions based on sample performances, is now recognised. Research in this area is set to continue well into the new century (Messick, 1994; McNamara, 1996).

[2] Unpredictability is not discussed in this paper. The 'open choice' principle upon which the unpredictability argument rests has been laid to rest, and discourse analysts have demonstrated that much language production is in fact highly predictable (Sinclair, 1987).

A legacy that is only just beginning to be investigated is the 'amalgam' of skills, or in its modern incarnation, the use of *integrative tasks* in language tests.[3] Early communicative language tests were certainly different in the degree to which integration was achieved, not in the sense of the term used by Oller (1979), but the deliberate thematic linking of test tasks, and the introduction of dependencies such as speaking based on a listening passage, where "a dependency must be accounted for" (Mislevy, 1995, p. 363). Admittedly, the "Communicative Use of English as a Foreign Language" (CUEFL, RSA) went particularly far in this, with no 'pure' scores for skills of language, but for task fulfilment (Morrow, 1977). Hargreaves (1987) reported that the main problems associated with this test (and similar tests) were "standardisation" (usually referred to as equating forms), and scoring. To these problems, Lewkowicz (1998) adds the fact that fewer tasks can be used in tests if they are not to become impractical, and tasks are more difficult and expensive to construct. Furthermore, reliability is associated with test length, and tests with integrated tasks may therefore need to be analysed in new and innovative ways. These problems currently seem insurmountable, but continued research into the reliable use of integrated tasks in performance tests will be one of the most important positive legacies of the communicative movement for the next few decades.

4.2. *Communicative tests are* authentic

1. *Purpose*: the test-taker must be able to recognise communicative purpose and be able to respond appropriately.
2. *Authenticity*: input and prompts in the language test should not be simplified for the learner.
3. *Context*: language will vary from context to context; some talk will be appropriate in one context and not another. The test-taker must therefore be tested on his/her ability to cope with the context of situation (physical environment, status of participants, the degree of formality, and expressing different attitudes), as well as the linguistic context.

Language for Specific Purpose (LSP), and more specifically EAP benefitted immediately and directly from the communicative movement. Carroll (1980, 1981, 1982) and Carroll and Hall (1985) took the principles of Munby (1978) and Wilkins (1976) to develop a framework for EAP test design, meeting the requirement of Morrow (1979) that test content should be tailored to learning needs, or purpose of communication. However, LSP testing has been plagued by the seeming impossibility of defining what is 'specific' to a particular communicative setting or purpose. Fulcher (1999) summarises the main findings of the field to date:

1. language knowledge (in its most general sense) accounts for most score variance in EAP tests;

[3] As in everything, integrated testing as currently being discussed is not new. Carroll (1961) first introduced the term, and Perren (1968) discussed the technical problems of integrated testing and dealing with item and task dependency in the "modern" sense.

2. grammar modules are better predictors of test total score than subject specific modules[4]; and
3. expert (content) judges cannot say what makes a text specific to their field.

The look and feel of EAP tests as they have evolved to date is likely to remain primarily because they 'look good' and encourage teachers and learners to focus on what is considered to be important in the classroom (Alderson, 1993, p. 216; Clapham, 1996, p. 201) rather than any empirical argument for their usefulness in testing (Clapham, 2000).

Perhaps more important is the role of *context* in its broadest sense. It has been recognised for some years that contextual variables in the testing situation impact upon discourse produced, and sometimes also upon test scores. This awareness has led language testers to conduct research into these variables, usually under the assumption that score variance which can be attributed to them constitutes systematic construct irrelevant variance. However, the view has been expressed that constructs as traditionally conceived by language testers cannot exist, because there is no such thing as competence, only variable performance and variable capacity (Tarone, 1998). Every change in context results in different performance. Accepting this extreme view from SLA research, Skehan (1987, p. 200) concluded that language testing was a matter of sampling, hence reducing the generalisability of test scores to the relationship between specific test task and its real-world counterpart. Tarone (p. 83) sees this as increasing validity while reducing reliability and generalisability. While the problems for language testing are clear (Fulcher, 1995), it is equally true that one of the most important tasks of the coming decade will be to identify the level of generalisability that is permitted in making inferences from scores across variable features of tasks. The most useful way of conceptualising this debate has been provided by Chapelle (1998), in which she distinguishes between (the new) behaviourism, trait theory, and interactionalism. While language testers are generally reluctant to abandon trait theory (and hence the treatment of contextual variance as error), some aspects of context may need to be defined as construct rather than error if they are found to be part of the meaning of test scores.

The question that remains is how far down this road, towards the extreme variationist position, language testing will move in theory. However, in practice, it is likely that the position adopted on the cline in the development of any specific test will be related to test purpose and the needs of score users, with the end of the cline in testing for extremely specific purposes (Douglas, 1998, p. 152, 2000a, b).

In early communicative texts, 'authenticity' meant the use of material in prompts that had not been written for non-native speakers of English, and a test could not be

[4] Davies (1978, p. 151) wrote: "What remains a convincing argument in favour of linguistic competence tests (both discrete point and integrative) is that grammar is at the core of language learning... Grammar is far more powerful in terms of generalisability than any other language feature. Therefore grammar may still be the most salient feature to test." Language testers like Clapham (2000), as well as second language acquisition researchers, are now returning to the study of grammar for precisely the reasons suggested by Davies. Bernhardt (1999, p. 4), for example, argues that, "...second language reading is principally dependent on grammatical ability in the second language".

communicative unless it was authentic. This was termed a "sterile argument" by Alderson (1981, p. 48). Modern performance tests that attempt to mirror some criterion situation in the external world are no more than role-plays or simulations, in which the learner is asked to 'imagine' that they are actually taking a patient's details in a hospital, giving students a mini-lecture, or engaging is a business negotiation. Language tests by their very nature are not mirrors of real life, but instruments constructed on the basis of a theory of the nature of language, of language use, and of language learning.

Widdowson (1983, p. 30) drew a distinction between the simple notion of 'authenticity' as conceived in early communicative writings, and:

> ...the communicative activity of the language user, to the engagement of interpretative procedures for making sense, even if these procedures are operating on and with textual data which are not authentic in the first [language produced by native speakers for a normal communicative purpose] sense.

In other words, the relationship between the learner and the task, how the learner deals with the task, and what we can learn about the learner as a result of doing the task, is what makes a task communicative.

In the most recent formulation of authenticity, Bachman and Palmer (1996) distinguish between authenticity (correspondence between test task characteristics and the characteristics of a target language use task) and interactiveness (abilities engaged by the task that are comparable to those engaged by the target language use situation). The degree to which the task characteristics match is the degree of task authenticity, and relates directly to construct validity through the ability to generalise from test task to target language use task.

This formulation combines previous definitions, relating the degree of authenticity achieved to the "perceptions" of authenticity by learners in specific situations (Lewkowicz, 2000, p. 49). In conjunction with a method for describing test and learner characteristics, the reformulation may aid research into construct validity in terms of how researchers hypothesise the construct operates over a range of different contexts in an interactionalist framework that expects significant score variation across contexts (Chapelle, 1998, pp. 41–42; Douglas, 1998).

However, Lewkowicz (1997, 2000) questions whether it is possible (or practical) to carry out comparisons of test tasks and target language use tasks using the new taxonomy of the Bachman and Palmer model, and also demonstrates that the perception of authenticity varies widely among test takers. Perhaps Chapelle (2000, p. 161) is right when she suggests that authenticity is a "folk concept", which may just be shorthand for 'all those bits of reality that we can't list but are simplified in models of test method/task facets'.

It is here where models of communicative language ability and test method (Canale and Swain, 1980, 1981; Bachman, 1990; Bachman and Palmer, 1996) are most useful in language testing research, whether the purpose is to isolate error variance or attempt to build task facets into construct definition. Such models may provide a framework for empirical research that will help define the constructs with

which language test developers work in the future. However, it should be remembered that models remain abstractions of perceptions of reality (Mislevy, 1995, p. 344), and as such are themselves constructs. Inferences are drawn from test scores to models of reality, and so the models remain in need of validation — or the estimation of the usefulness of the model for the research for which it is being used (Fulcher, 1998b).

4.3. *Communicative tests are scored on* real-life outcomes

> *Behaviour-based*: the only real criterion of success in a language test is the behavioural outcome, or whether the learner was able to achieve the intended communicative effect.

When students are asked to perform a task, it is essential to state what it is that the raters are measuring. The "behavioural outcomes" of Morrow (1979) are difficult to isolate, and even more difficult to specify (Fulcher, 1999, pp. 224–225). And it is also possible for learners to successfully complete acts of communication with little or no knowledge of a language. In performance tests it is through the rating scale that what is being tested is defined. In fact, more research has been carried out into rating scales than any other aspect of tests of speaking (Fulcher, 1998a), and it is perhaps here where the early British communicative movement has had no lasting legacy for modern testing practice. Until recently, the template for rating scale design was the original FSI, and rating scales designed for British tests have not fared well from critics (Fulcher, 1987; Wood, 1990).

Most rating scale design has been a-theoretical, relying purely on arm-chair notions of language development and structure. Only recently have empirical methods been employed in the construction and validation of rating scales (Fulcher, 1993, 1996; North, 1996; North and Schneider, 1998; Upshur and Turner, 1995). Research in this area is now growing (Turner, 1998), and as rating scales of more components of models of communicative language ability are designed, operationalised and used in research, language testers and SLA researchers will learn more about language and how it is used across a range of tasks and target language use situations by speakers with different characteristics.

5. Conclusion

Much current research in language testing may have developed despite the British communicative testing movement. The concerns with language use that have generated the questions that are now being addressed already existed in the United States. However, the ideology associated with the early British communicative movement has had a pervasive influence on the ethos of teaching, learning and testing around the world. It would be difficult to market a new large-scale test that did not claim to be 'communicative' — whatever the term may mean for different users. The positive legacy of the movement as a whole can be seen in all the research that concerns itself

with more careful definitions of task, of context, and the relationship between test task and the target language use situation. This is especially the case in testing language for specific purposes. Yet, this research is being done with a rigour that was rejected by the British communicative testing movement, involving the use of new statistical tools and the development of sophisticated conceptual paradigms. We may even see the development of high quality tests conforming to the standards expected in the United States (APA, 1999) with the attractive and creative content of British tests.[5] Whatever specific benefits there may be, the unified concept of validity, a generally accepted theoretical framework for research in educational assessment, and growing interest in the ethics of language testing[6] should lead to a decade of cooperative research that will bring significant advances in language testing theory and practice.

Acknowledgements

My thanks are due to Bob Hill, whose observations on the vagueness of popular terminology in language testing led me to consider the legacy of 'communicative testing' at the start of a new decade. And to Fred Davidson who provided constructive criticism on the first draft of this paper. Responsibility for any errors, and the views expressed, remains mine.

References

Adams, M.L., Frith, J.R., 1979. Testing Kit: French and Spanish. Department of State, Foreign Service Institute, Washington D.C.

Alderson, J.C., 1981. Reaction to the Morrow Paper (3). In: Alderson, J.C., Hughes, A. (Eds.), Issues in Language Testing. The British Council, London, pp. 45–54.

Alderson, J.C., 1993. The relationship between grammar and reading in an English for academic purposes test battery. In: Douglas, D., Chapelle, C. (Eds.), A New Decade in Language Testing Research. TESOL Publications, Washington DC, pp. 203–219.

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

ALTE, 1994. The ALTE Code of Practice. Cambridge Local Examinations Syndicate, Cambridge.

APA (American Educational Research Association, American Psychological Association, National Council on Measurement in Education), 1999. Standards for Educational and Psychological Testing. AERA, Washington DC.

Bachman, L.F., 1976. Fundamental Considerations in Language Testing. Oxford University Press, Oxford.

Bachman, L.F., Davidson, F., 1990. The Cambridge-TOEFL comparability study: an example of the cross-national comparison of language tests. In: de Jong, H.A.L. (Ed.), Standardization in Language Testing. AILA Review, Amsterdam, pp. 24–45.

[5] The TOEFL 2000 project may lead to the first language test that achieves the integration of high technical quality, innovative task types, and construct frameworks that guide test design (Jamieson et al., 2000).

[6] See Messick (1981, 1984, 1989a, b, 1994).

Bachman, L.F., Palmer, A.S., 1996. Language Testing in Practice. Oxford University Press, Oxford.
Barnwell, D.P., 1996. A History of Foreign Language Testing in the United States: From its Beginnings to the Present Day. Bilingual Press, Arizona.
Bernhardt, E., 1999. If reading is reader-based, can there be a computer adaptive test of reading? In: Chalhoub-Deville, M. (Ed.), Issues in Computer-adaptive Testing of Reading Proficiency. Cambridge University Press, Cambridge, pp. 1–10.
Canale, M., Swain, M., 1980. Theoretical bases of communicative approaches to second language teaching and testing. Applied Linguistics 1 (1), 1–47.
Canale, M., Swain, M., 1981. A theoretical framework for communicative competence. In: Palmer, A.S., Groot, P.J.M., Trosper, G.A. (Eds.), The Construct Validation of Tests of Communicative Competence. TESOL Publications, Washington DC, pp. 31–35.
Carroll, J.B., 1961. Fundamental considerations in testing for English language proficiency of foreign students. In: Allen, H.B. (Ed.), Teaching English as a Second Language. McGraw Hill, New York, pp. 364–372.
Carroll, J.B., 1967. The foreign language attainments of language majors in the senior year: a survey conducted in U.S. colleges and universities. Foreign Language Annals 1 (2), 131–151.
Carroll, B.J., 1980. Testing Communicative Performance: An Interim Study. Pergamon, Exeter.
Carroll, B.J., 1981. Specifications for an English Language Testing Service. In: Alderson, J.C., Hughes, A. (Eds.), Issues in Language Testing. The British Council, London, pp. 68–110.
Carroll, B.J., 1982. Language testing: is there another way? In: Heaton, J.B. (Ed.), Language Testing. Modern English Publications, London, pp. 1–10.
Carroll, B.J., Hall, P.J., 1985. Make Your Own Language Tests: A Practical Guide to Writing Language Performance Tests. Pergamon, Oxford.
Chapelle, C., 1998. Construct definition and validity inquiry in SLA research. In: Bachman, L.F., Cohen, A.D. (Eds.), Interfaces Between Second Language Acquisition and Language Testing Research. Cambridge University Press, Cambridge, pp. 32–70.
Chapelle, C., 2000. From reading theory to testing practice. In: Chalhoub-Deville, M. (Ed.), Issues in Computer-adaptive Testing of Reading Proficiency. Studies in Language Testing, Vol. 10. Cambridge University Press, Cambridge, pp. 150–166.
Clapham, C., 1996. The Development of IELTS. A Study of the Effect of Background Knowledge on Reading Comprehension. Cambridge University Press, Cambridge.
Davies, A., 1978. Language testing. In: Language Teaching and Linguistics Abstracts. Vol. 11, 3 & 4, reprinted in Kinsella, V. (Ed.) (1982) Surveys 1: Eight State-of-the-art Articles on Key Areas in Language Teaching. Cambridge University Press, Cambridge, pp. 127–159.
Douglas, D., 1998. Testing methods in context-based SL research. In: Bachman, L.F., Cohen, A.D. (Eds.), Interfaces Between Second Language Acquisition and Language Testing Research. Cambridge University Press, Cambridge, pp. 141–155.
Douglas, D., 2000a. Assessing Languages for Specific Purposes. Cambridge University Press, Cambridge.
Douglas, D., 2000b. Testing for specific purposes. In: Fulcher, G., Thrasher, R. Video FAQs: Introducing Topics in Language Testing. Available at: http://www.surrey.ac.uk/ELI/ilta/faqs/main.html
Edgeworth, F.Y., 1890. The element of chance in competitive examinations. Journal of the Royal Statistical Society 49 (1), 644–663.
Fulcher, G., 1987. Tests of oral performance: the need for data-based criteria. English Language Teaching Journal 14 (4), 287–291.
Fulcher, G., 1993. The construction and validation of rating scales for oral tests in English as a Foreign Language. Unpublished PhD thesis, University of Lancaster, UK.
Fulcher, G., 1995. Variable competence in second language acquisition: a problem for research methodology. System 23 (1), 25–33.
Fulcher, G., 1996. Does thick description lead to smart tests? A data-based approach to rating scale construction. Language Testing 13 (2), 208–238.
Fulcher, G., 1998a. The testing of speaking in a second language. In: Clapham, C., Corson, D. (Eds.), Language Testing and Assessment. Encyclopedia of Language and Education. Vol. 7. Kluwer Academic Publishers, Dordrecht, pp. 75–86.

Fulcher, G., 1998b. Widdowson's model of communicative competence and the testing of reading: an exploratory study. System 26, 281–302.
Fulcher, G., 1999. Assessment in English for Academic Purposes: putting content validity in its place. Applied Linguistics 20 (2), 221–236.
Fulcher, G., 2000. Computers in language testing. In: Brett, P., Motteram, G. (Eds.), Computers in Language Teaching. IATEFL Publications, Manchester, pp. 97–111.
Hamp-Lyons, L., 1987. Cambridge First Certificate in English. In: Alderson, J.C., Krahnke, K.J., Stansfield, C.W. (Eds.), Reviews of English Language Proficiency Tests. TESOL Publications, Washington DC, pp. 18–19.
Hargreaves, P., 1987. Royal Society of Arts: examinations in the communicative use of English as a Foreign Language. In: Alderson, J.C., Krahnke, K.J., Stansfield, C.W. (Eds.), Reviews of English Language Proficiency Tests. TESOL Publications, Washington DC, pp. 32–34.
Harrison, A., 1983. Communicative testing: jam tomorrow? In: Hughes, A., Porter, D. (Eds.), Current Developments in Language Testing. Academic Press, London, pp. 77–85.
Hayden, P.M., 1920. Experience with oral examinations in modern languages. Modern Language Journal 5, 87–92.
Jamieson, J., Jones, S., Kirsch, I., Mosthenthal, P., Taylor, C., 2000. TOEFL 2000 framework: a working paper. ETS: TOEFL Monograph Series 16, Princeton NJ.
Kaulfers, W., 1944. War-time developments in modern language achievement tests. Modern Language Journal 28, 136–150.
Lado, R., 1961. Language Testing: The Construction and Use of Foreign Language Tests. Longman, London.
Lewkowicz, J.A., 1997. Authenticity for whom? Does authenticity really matter? In: Huhta, A., Kohonen, V., Lurki-Suonio, L., Luoma, S. (Eds.), Current Developments and Alternatives in Language Assessment. Jyvaskyla University, Finland, pp. 165–184.
Lewkowicz, J.A., 1998. Integrated testing. In: Clapham, C., Corson, D. (Eds.), Language Testing and Assessment. Encyclopedia of Language and Education, Dordrecht: Kluwer Academic Publishers, Vol. 7, pp. 121–130.
Lewkowicz, J.A., 2000. Authenticity in language testing: some outstanding questions. Language Testing 17 (1), 43–64.
Liskin-Gasparro, J.E., 1984. The ACTFL proficiency guidelines: gateway to testing and curriculum. Foreign Language Annals 17 (5), 475–489.
Lowe, P., 1976. Handbook on Question Types and the Use in the LS Oral Proficiency Tests. CIA Language School, Washington DC.
Lowe, P., 1983. The ILR oral interview: origins, applications, pitfalls, and implications. Die Unterrichtspraxis 16, 230–244.
Lundeberg, O.K., 1929. Recent developments in audition-speech tests. Modern Language Journal 14, 193–202.
McNamara, T., 1996. Measuring Second Language Performance. Longman, London.
Mercier, L., 1933. Diverging trends in modern foreign language teaching and their possible reconciliation. French Review 6 (3), 370–386.
Messick, S., 1981. Evidence and ethics in the evaluation of tests. Educational Researcher 10 (9), 9–20.
Messick, S., 1984. Assessment in context: appraising student performance in relation to instructional quality. Educational Researcher 13 (3), 3–8.
Messick, S., 1989a. Validity. In: Linn, R.L. (Ed.), Educational Measurement. Macmillan, New York, pp. 13–103.
Messick, S., 1989b. Meaning and values in test validation: the science and ethics of assessment. Educational Researcher 18 (2), 5–11.
Messick, S., 1994. The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher 23 (2), 13–23.
Mislevy, R.J., 1995. Test theory and language learning assessment. Language Testing 12 (3), 341–369.
Morrow, K., 1977. Techniques of Evaluation for a National Syllabus. Royal Society of Arts, London.

Morrow, K., 1979. Communicative language testing: revolution of evolution? In: Brumfit, C.K., Johnson, K. (Eds.), The Communicative Approach to Language Teaching. Oxford University Press, Oxford, pp. 143–159.
Morrow, K., 1982. Testing spoken language. In: Heaton, J.B. (Ed.), Language Testing. Modern English Publications, London, pp. 56–58.
Morrow, K., 1991. Evaluating communicative tests. In: Anivan, S. (Ed.), Current Developments in Language Testing. RELC, Singapore, pp. 111–118.
Munby, J.L., 1978. Communicative Syllabus Design. Cambridge University Press, Cambridge.
North, B., 1996. The development of a common framework scale of descriptors of language proficiency based on a theory of measurement. Unpublished PhD thesis, Thames Valley University.
North, B., Schneider, G., 1998. Scaling descriptors for language proficiency scales. Language Testing 15 (2), 217–262.
Oller, J., 1979. Language tests at School. London, Longman.
Perren, G.E., 1968. Testing spoken language: some unsolved problems. In: Davies, A. (Ed.), Language Testing Symposium: A Psycholinguistic Approach. Oxford University Press, Oxford, pp. 107–116.
Sinclair, J. McH, 1987. Collocation: a progress report. In: Steele, R., Treadgold, T. (Eds.), Essays in Honour of Michael Halliday. John Benjamins, Amsterdam, pp. 319–331.
Skehan, P., 1987. Variability and language testing. In: Ellis, R. (Ed.), Second Language Acquisition in Context. Prentice Hall, Hemel Hempstead, pp. 195–206.
Sollenberger, H.E., 1978. Development and current use of the FSI oral interview test. In: Clark, J.L.D. (Ed.), Direct Testing of Speaking Proficiency: Theory and Application. Educational Testing Service, Princeton, NJ, pp. 1–12.
Spolsky, B., 1976. Language Testing: Art of Science? Paper read at the 4th International Congress of Applied Linguistics. Stuttgart, Germany.
Spolsky, B., 1995. Measured Words. Oxford University Press, Oxford.
Tarone, E., 1998. Research on interlanguage variation: implications for language testing. In: Bachman, L.F., Cohen, A.D. (Eds.), Interfaces Between Second Language Acquisition and Language Testing Research. University of Cambridge Press, Cambridge, pp. 71–89.
Turner, J., 1998. Assessing Speaking. Annual Review of Applied Linguistics. Cambridge University Press, Cambridge, 192-207.
Underhill, N., 1982. The great reliability validity trade-off: problems in assessing the productive skills. In: Heaton, J.B. (Ed.), Language Testing. Modern English Publications, London, pp. 17–23.
Underhill, N., 1987. Testing Spoken Language. Cambridge University Press, Cambridge.
Upshur, J.A., Turner, C.E., 1995. Constructing rating scales for second language tests. English Language Teaching Journal 49, 3–12.
Widdowson, H.G., 1983. Learning Purpose and Language Use. Oxford University Press, Oxford.
Wilkins, D.A., 1976. Notional Syllabuses. Oxford University Press, Oxford.
Wood, B.D., 1927. New York Experiments with New-Type Modern Language Tests. Macmillan, New York.
Wood, R., 1990. Assessment and Testing: A Survey of Research. Cambridge University Press, Cambridge.