STATE OF THE ART IN *LANGUAGE TESTING:* TEN YEARS OF HISTORY OR HOW TO DEAL WITH ASSESSMENT

HONESTO HERRERA SOLER
UNIVERSIDAD COMPLUTENSE

1. INTRODUCTION

More than a decade has passed since the *Language Testing* journal (*LT*) was first published. Now it is time to look back and gauge whether it has fulfilled its aims in this interval and whether it has successfully filled the research gap it was intended to fill. In the first issue, the then editors set out their aims as follows:

> This new journal has come into being as a forum devoted exclusively to the issues which concern those involved with, or simply interested in, the assessment of language ability in one form or another . . . the field covered will be a broad one. The journal will consist of contributions from second or foreign language testing, mother tongue testing and the assessment of language disability.... (Editorial board, 1984: 1)

Miscelánea: A Journal of English and American Studies 18 (1997): 115-137.

The editors went on to outline their view of the role of the journal: preference would be given to articles on theoretical issues based on empirical research or otherwise throwing some light on the field of language testing. Time has passed and not much change is to be appreciated in these aims and purposes in the intervening years, though now, once *LT* has come of age and has established its identity, it can be said that its fields of study have widened.

Our purpose in this article is to go through these ten years of *LT* issues, since the underlying philosophy, purpose and range of topics of the first issues of the beginning of the second decade are quite similar to the first one, to review the research done on the different features of testing in second language acquisition, to identify the main topics and to analyse to what extent this journal has been an answer to the needs of the testing community. Aware as we are of the doubts, fears and misgivings most teachers with an arts background have when faced with correlation coefficients, standard deviations, etc., we have attempted to deal in greater detail with the contributions in which statistics are applied to linguistics and have highlighted the efforts made to update the different testing techniques during this decade. A substantial number of the problems we have to cope with in our classroom every day have been dealt with in *LT* pages, and not a few answers to our difficulties can be found there.

In its short history, *LT* has published articles which could be considered an end in themselves, that is, dealing with a particular feature, process or method of testing. Other articles, rather than being concerned with the specific testing technique in itself, study some of the hundred or so issues which may crop up either in the language teaching or language learning field. In the latter case, testing would be just a tool at the service of a learning and teaching hypothesis. The common core of the topics for future articles proposed by the editorial board in this period was the assessment of language ability. The most frequently dealt with topics were acquisition of a second language, methods, testing strategies and certain issues in linguistics fields; not an issue went by without an article on the Item Response Theory (IRT) and on the Testing of English as a Foreign Language (TOEFL) examination, the two main lines of research in *LT*. The articles published in the first two issues provide the route map to be followed and developed in the subsequent ones.

2. 1984: A KEY MOMENT IN THE TEN-YEAR HISTORY OF *LT*

The 1984 issue of *LT* lays down the main areas of interest which were to be developed in subsequent issues. Thus, the range of topics covered that year—validity and reliability, criterion-referenced measurement versus norm-referenced measurement, the unitary competence hypothesis, specific and technical issues on testing, the introduction to the Item Response Theory (IRT) and the articles on Testing of English as a Foreign Language (TOEFL)—may be considered to be the main threads of the story of *LT*.

Reliability—the extent to which the results in a test can be considered consistent and stable—and validity—the degree to which a test measures what it claims to be measuring—in testing are the first two features to be dealt with in depth, in articles by Krzanowsky and Woods (1984) and Davies (1984). The first article, a good introduction to the use to which a linguist may need to put statistics, deals with some simple analysis of variance (ANOVA) models which can be used to define and estimate reliability coefficients: the Spearman-Brown, concerned with within forms and between forms reliability, the Cronbach's alpha, where split-parts estimates of reliability are considered, or the Kuder-Richardson formula 20, a limiting case of Cronbach's alpha when k becomes equal to the number of items in the test, and each score then simply takes the value *1* (for correct) and *0* (for incorrect). Davies discusses the process of concurrent and predictive validation for the English Proficiency Test Battery (EPTB), English Language Battery (ELBA), English Language Testing Service (ELTS), using students' grades or examination results and the teachers' or tutors' estimates as criteria for validity purposes. Hudson and Lynch (1984) tackle the reliability and validity issues as well, but this time they focus their research on criterion-referenced measurement (CRM) versus norm-referenced measurement (NRM) methods,[1] that is, they are concerned with the way the results are interpreted.

It is assumed that the reader is familiar with the IRT,[2] and the article presents research done from the linguistic point of view on some fundamentals

of this theory: the IRT versus the CTT (Classical Test Theory). The IRT is an attempt to overcome the conflict we come across in our traditional tests: Is the test too easy / too difficult for this specific group of students or, on the contrary, is the level of these students too high / too low for this test? The aim of this technique is to help us to build unbiased tests in which all individuals having the same basic ability are equally likely to get the item correct, regardless of subgroup membership or the testing technique where the item might appear.

The advantages of the IRT, a technique which tries to attenuate discrepancies between student ability and item difficulty, over the CTT are discussed in the contribution of Perkins and Miller (1984). According to their research, the IRT detects more misfitting and weak items than the classical test theory indices. Henning (1984) goes a step further when he studies the advantages of latent trait—the unobservable ability—measurement in language testing. This article presents not only the Rasch Model latent trait procedures as an alternative to classical measurement theory but also the analytical procedures: the Rasch one-parameter logistic model, and the Birnbaum two/three parameter logistic model. The Rasch model is concerned with a single ability-difficulty parameter, while the others incorporate additional parameters of discriminability and guessing. The IRT means that we must cope with probabilistic models since they try to evaluate items and persons, not only in classical terms of difficulty, ability, variance and discriminability, but also in terms of quantifiable deviations from predicted response patterns.

Henning was the first to deal with the IRT from a theoretical point of view in the second issue of 1984. Over the next ten years he was to become the main contributor on IRT. Nevertheless, firm supporter though he is of this theory, we read at the end of his article that in spite of the supposed advantage of the IRT he suggests that the classical measurements should not be abandoned but should be supplemented through the informational advantage of latent trait and item response theory.

The rest of the articles in the 1984 issue discuss either specific theoretical issues (students' reaction to tests, the possibility of characterizing language impairment, etc.) or empirical studies. Thus, Skehan refers to Oller's unitary competence hypothesis, an attempt to demonstrate that one underlying competence, a general factor, accounts for language performance. Other articles are concerned with testing techniques: Shohamy discusses multiple

choice versus open-ended questions, Bensoussan studies cloze tests versus multiple choice, and Klein-Braley and Raatz apply themselves to overcoming the cloze drawbacks with their C-Test.

3. THE IRT, A THREAD RUNNING THROUGH *LT*

The 1984 contributions accurately forecast the main areas of interest in the following years. There is no year in which in one way or another the IRT is not considered. It is a thread that runs all through the journal. The range of approaches to this topic is very wide, because everybody endeavours to test their hypothesis with the IRT. Nevertheless, it would be wrong to conclude that this technique is a panacea for the mismatch between student ability and item difficulty or the discrimination or guessing problems, corresponding to one-parameter, two-parameter or three-parameter logistic models respectively. There is still much to be done despite the IRT contributions, and the testing community as a whole should be aware not only of its advantages but also of its shortcomings. The effort involved for those with an Arts background in understanding the contribution of this theory to testing may well be worthwhile, whether we are involved in General English or English for Specific Purposes.

The orientation of research on this theory tends to be either theoretical oriented or practical. Among the former type are contributions in which the classical methods are challenged and others in which the meaning of the three models is discussed. There are also studies on the partial credit model,[3] the issue of unidimensionality, and two different approaches to IRT. The practically oriented contributions deal mainly with reading and listening comprehension tests.

a) Theoretical issues

The IRT, though considered an alternative to the traditional methods, is reduced to the Rasch Model, one-parameter, in the Woods and Baker (1985) contribution. The IRT is presented as a tool to measure on the same scale the ability of the subjects and the difficulty of the items. In the end, the value of the Rasch analysis, according to Woods and Baker, will depend on how much information testers can extract from it, information which using classical methods could not be obtained at all or only with difficulty.

Henning et al. (1985) go further in the analysis of the IRT. They attempt to investigate the robustness and applicability of the Rasch Model for use in language proficiency tests that consist of batteries and subtests in a variety of skill areas. In a second study, Henning tries to demonstrate the utility of Rasch Model scalar analysis when applied to self-ratings of ability / difficulty associated with component skills of English as a second language. Hudson (1993) investigates relationships among the IRT one-parameter fit statistics, the two-parameter slope and traditional biserial correlations in terms of the role these indices play in criterion-referenced language test construction.

The Partial Credit model, an extension of the simple Rasch dichotomous model (Rasch 1960, 1980) is discussed by Adams et al. (1987) as an alternative to the classical test theory. This model allows for the scoring of items in any number of ordered categories as the basis for the construction and analysis of an oral interview test. It is also demonstrated in Tomlinson et al. (1988) that item forms, Rasch Partial Credit Model, can be developed for verbal tasks based on grammatical or structural organizing framework. Finally, Pollit and Hutchinson (1987) describe the use of the partial credit form of the Rasch model in the analysis and calibration of a set of writing tasks. For this kind of analysis it is necessary that the tasks be carefully controlled and that the assessment scales and criteria be adapted to suit the specific demands of each task. They conclude that with the availability of the partial credit version of the Rasch model it is now possible to analyse any form of assessment which produces numerical outcomes.

The person-characteristic function (PCF), an opposite approach to the IRT, is developed by John Carroll (1986). It consists in relating the probability of an individual's passing an item to the difficulty of the item, over items, i. e. it depends on the item's difficulty, whereas IRT is concerned with the probability of passing as a function of ability, over individuals, i. e. it depends on the individual's ability. With the item information functions (IIF), Hudson (1993) examines the relationship of three item discrimination indices and the biserial correlation to IRT in order to provide testers with information which will be useful in contexts in which IRT analysis is inappropriate.

In addition to the models and the different approaches to the IRT, Henning et al. (1985) were concerned with unidimensionality. Their study was designed to test the effects of violation of the unidimensionality assump-

tion on Rasch Model estimates of item difficulty and person ability. The results clearly suggested that violations of item unidimensionality produced distorted estimates of item difficulty. The Bejar method was found to be sensitive to such distortions, and results of applying the Bejar Method along with internal consistency estimation and principal components analysis were mutually confirmatory. Henning (1989) further discusses this topic. In this article it is argued that local independence, unidimensionality, and noninvasiveness are important but distinct concepts that may, but need not necessarily, overlap.

b) Applicability

The purpose of the study by Choi and Bachman (1992) was to examine the appropriateness and adequacy of the 1-, 2- and 3-parameter logistic IRT models for analysing data from two EFL proficiency tests. Theunissen's (1987) study of applicability refers to reading comprehension tests, and so does Boldt's (1989), which deals with the use of IRT method—this time taking into account the population—to study nonlinguistic issues on testing: cultural background, native speakers as raters, speed in dealing with grammatical reasoning, constraints on cloze testing and cognitive strategies in reading comprehension. He discusses the possibility of computerizing the test design, the calibration of items, the Rasch Model and the concept of test reliability replaced in item response theory by the vastly superior concept of test information. Not only reading comprehension tests but also listening comprehension tests are discussed, and de Jon and Glas (1987) have recourse to IRT for their validation.

Beyond the General English Tests concern, we read the McNamara (1990) discussion on the role of Rasch Model IRT in the validation of two sub-tests of the Occupational English Test. He argues for the usefulness of IRT as a tool in the implications of the empirical analysis presented for the validity of communicative language tests involving the skills of speaking and writing. McNamara (1991) is concerned with another skill: a listening test in ESP. The study confirms the appropriateness of IRT approaches to the analysis of data from a ESP test. The useful role of Rasch IRT in the investigation of the content and construct validity of language tests is also confirmed.

4. TOEFL: A SECOND THREAD FOR THE TEXTURE OF *LT*

The articles on the standardised Test of English as a Foreign Language, like those on IRT, are like the weft and woof of *LT*'s texture. Whether it is a normal or a special issue, articles can always be found either on the item response theory or on this standardised test: the researcher will consider the topic suggested by the editors from the IRT or TOEFL perspective respectively. The studies done on this test are less concerned with theory than with practice. Researchers try either to improve some of the TOEFL batteries, compare it with other tests or refer their studies on reading and listening comprehension and on written English to TOEFL.

Although Spolsky's (1990) is not the first contribution on this topic it is the first from the thematic point of view. As when he studied the three phases of testing he is also concerned here with the prehistory of this test. He presents a report on the origins of the TOEFL and its development, together with the main comments on this battery of tests in the conference held in Washington on May 11-12, 1961. His target is to gain an understanding of how developments in language testing theory are blended with the requirements and possibilities of real life implementations.

Among the contributions, whose purpose is to improve this standardised test, we read Stansfield and Ross (1988) on the one hand and Boldt (1989, 1992) on the other. The former deals with the validity and reliability of the Test of Written English (TWE) commissioned by the TOEFL research committee, where concurrent, predictive, content and face validity, and reliability are discussed. The latter copes with the latent structure analysis of the Test of English as a Foreign Language. Equating[4] studies support the use of IRT methods for TOEFL. This is done separately for each of three sections of TOEFL: Listening Comprehension, Structure and Written Expression, and Reading Comprehension and Vocabulary. The model assumes that a single latent proficiency variable underlies item performance, but the TOEFL candidate population is diverse, perhaps containing many groups with separate latent variables. Later, Boldt (1992) carries out a crossvalidation study in which a proportional item response curve (PIRC) is used to predict item scores of selected examinees on selected items.

A good battery of tests needs constant innovations which take into account social, cultural and linguistic changes if it is to be considered a highly successful test among tests takers and institutions. Thus, the Princeton-based Educational Testing Service encourages all sort of studies that may lead to check its advantages as well as its drawbacks. Consequently, it is not surprising to see that most of the research done on this major modern test of English as a foreign language in these pages has been through the comparative method. Bachman et al. (1988) compare two EFL proficiency test batteries. The Cambridge-TOEFL analysis is based on the abilities measured by the two tests. Bachman et al. focus on the qualitative analysis—the description of the abilities that appear to be measured and of the tasks required of the test takers—rather than on the quantitative examination of test performance. Ryan and Bachman (1992) carry out comparisons and examine the extent to which items from two widely-used EFL proficiency test TOEFL and FCE function differently for test-takers of equal ability from different native language and curricular backgrounds. De Mauro (1992) examines the relationships among the Test of Spoken English (TSE), the Test of Written English (TWE), and TOEFL. The multivariate prediction of each of these tests from the scores on the others is very accurate. Finally, Hale et al. (1989) in their research of four categories of multiple-choice (MC) cloze items take the TOEFL as a point of reference.

Test takers are also taken into account. Powers (1986) studies the listening comprehension section of the TOEFL. In his research, differences between native and non-native speakers in relation to each of the listening skills and the appropriateness of general or specific tasks for evaluating listening skills are discussed. Hale (1988) hypothesizes that the student's academic discipline will interact with the text content in determining performance on the reading passages of the TOEFL. Freedle and Kostis (1993) also deal with reading comprehension but at the item level. They set out to examine whether text and text-by-item interaction variables play a significant role in predicting item difficulty.

5. AUTHENTICITY

Like IRT and TOEFL, another issue which crops up frequently is authenticity, one of the main building blocks of this journal since, in discussing communicative language testing, questions relating to the issue of authenticity—tasks, texts, content and setting—form a recurrent theme. Aware as the *LT* editorial board was of the general principle—the greater the similarity of a test to the performance to be assessed or predicted, the greater the likelihood that the test will a) be predictive of future performance and b) accepted by the test users and test takers—they focused their attention on this topic in the first 1985 issue and questions relating to authenticity were studied in several articles over the years.

Assuming that the success of communication can only be measured by the degree to which the meaning intended in the mind of the speaker is generated in the mind of the listener, Seliger (1985) draws attention to the problem of meaning. His analysis leads him to consider the types of inequality to be overcome: the adult-child interaction and the native speaker versus the second language learner. Therefore, the utilization of extralinguistic information in developing inferences or hypotheses to resolve conditions of incomprehension at linguistic and pragmatic level are required. Seliger discusses the possibility that language testing should develop tests that focused not on the product but on the successful implementation of the processes. Perspective which would constrain testing to criterion-referenced tests.

Further studies on pragmatics and the testing of communicative competence are carried out by Olshtain and Blum-Kulka (1985). They argue that while most areas of grammatical competence can and have already been translated into operational, dichotomous testing items, the complexity in translating components of communicative competence into testing items stems from the lack of sufficient systematic studies in native language use.

Rather than with interaction or communicative competence problems in testing, Spolsky (1985) is concerned with the limits of authenticity in language testing, since any language test is by its very nature inauthentic. The test taker is being asked not to answer a question giving information but to display knowledge or skill. Shohamy and Reves (1985) on the topic of authenticity distinguish between the language of authentic tests and real life language. They argue that if we insist on eliciting authentic real-life language we should adopt an ethnographic approach. This approach to authenticity, in

which the boundaries are not well-defined since we cannot leave aside the pragmatic or ethnographic influence on linguistics, entails some major deficiencies such as the lack of measurement, statistical analysis and limited empirical evidence.

6. SELF-ASSESSMENT

The tester and the testee go together in self-assessment. Literature on this topic was quite new at a time in which students had tools, such as computers, data banks of items, etc., to assess their proficiency in a way that they could not have dreamed of a few decades ago. The fact that technology was so advanced and had become quite fashionable could have been a good reason for the editorial board to propose self-assessment as the central topic of some of its first issues. Before this publication appeared some previous research had been carried out in this field: Oscarson (1977) self-assessment can yield quite informative results and Von Telek (1982) found correlations between self-assessment and follow-up tests. It is worth mentioning Le Blanch and Painchaud's (1985) investigations on the usefulness of self-assessment as a second language placement instrument, and Davidson and Henning's (1985) conclusion that little confidence should be placed in the specific student self-rating they examined.

The beliefs which underlie the idea of self-directed learning—where the learner is learning to do something rather than about something—and consequently self-assessment, and the reasons which can be adduced in support of those beliefs, are examined by Houghton and Dickinson (1988), who put forward a scheme in which they try to reconcile the tensions between self-assessment and institutional assessment leading to certification.

Most of the articles in the first issue of 1989 are devoted to self-assessment. Oscarson outlines the justification for adopting self-assessment principles in language teaching and learning, since he thinks that it should be oriented to formative purposes rather than purposes such as selection,

grading and certification. A detailed comparison is carried out between a test of Dutch as a second language for use in language courses for adult learners, and a parallel version of that test in self-assessment format in Janseen (1989). Meanwhile, Bachman and Palmer investigate the structure of an experimental self-rating test of communicative language ability through the use of multi-trait multimethod (MTMM) design and confirmatory factor analysis (CFA). The language abilities intended to be measured comprised three main traits: grammatical competence, pragmatic competence and sociolinguistic competence. The reliabilities obtained were much higher than had been expected, and all the self-rating measurements had strong loadings on a general factor. Measurements of grammatical competence appear to be better indicators of this trait than measurements of pragmatic and sociolinguistic competence. Finally, the role of response effects—the tendency to respond to factors other than item content—is investigated. Results in Heilenman (1990) indicate that both an acquiescence effect—a tendency to respond positively regardless of item content—and overestimation were present and more evident in less experienced learners.

7. EXTRALINGUISTIC FACTORS

a) Cultural background and affective reactions

The primary aim of Zeidner's (1987) study was to test for ethnic, sex, and age bias in the predictive validity of English language aptitude test scores. Overall, the results of this research are in line with the bulk of previous studies on cultural bias, reporting a slight degree of intercept bias when cognitive indices are used in predicting first year college grade. Chihara et al. (1989) also discuss background and culture as factors in EFL reading comprehension presenting two versions of clozes, one original and the other modified introducing mainly proper names. Not only materials but also the influence of the different types of tests on students are considered as well. Zeidner and Bensoussan (1988) analyse college students' attitudes towards written versus oral tests of English as a Foreign Language; their data are based on students' interests and preferences. No meaningful relationship is observed. Two years later, Bradshaw also takes the issue of the test-takers' attitudes to a placement test. She concludes that some sort of feedback from

test consumers should be included together with issues of content and construct validity, statistical reliability and practicality when we prepare a test.

Not only attitudes but also feelings and conditions in which tests are taken are studied in the affective reactions of native Brazilian students to different oral EFL test formats in an achievement testing situation. Scott (1986) assesses factors like format, length, time constraint, testing environment, familiarity with test format, perceptions of test validity, and student anxiety.

b) Strategies

Nevo (1989) reports research whose purpose was to study the processing of reading comprehension tests and to ascertain the cognitive strategies. In his test-wiseness scale, Allan (1992) goes further and finds that students are differentially skilled in test taking and that the scores of some learners may be influenced by skills which are not the focus of the test, thus invalidating their results. Amer (1993) investigates the effect of teaching a test-taking strategy to EFL students on their performance on EFL test. He considers the following components of a test-taking strategy: to read the instructions carefully, to schedule their time appropriately, to make use of clue words in the questions, to delay answering difficult questions, and to review their work in order to check their answers. Components which were summarised in Carman and Adams (1972) "scorer acronym":

S - Schedule your time
C - Clue words
O - Omit difficult questions
R - Read carefully
E - Estimate your answer
R - Review your work

8. TESTING TECHNIQUES

A glance at the inside cover of any issue will confirm that it is assumed that theoretical issues and empirical research must go together, since any hypoth-

esis, if it is to be tested, needs some sort of data base. There is a tendency to measure everything. Apart from IRT, which has been widely commented on, there are all kinds of testing techniques which may help to quantify the central feature of any study. The articles gathered under this heading are more concerned with the "how"—the technique itself— than with the "what"—any aspect of language teaching or learning processes. Although the use of these testing techniques is usually a means, in some cases there is a tendency to consider these techniques as ends in themselves. The articles chosen for comment range from issues such as controversy, reliability and validity of the cloze and multiple choice to the different versions of these techniques.

a) Traditional clozes

Controversy on cloze usefulness is to be found in Lado (1986). He responds to Oller and Conrad's (1971) point of view—they consider the cloze method extremely useful in the placement of non-native speakers of English and in the diagnosis of their special language problems— whereas Lado does not share that perspective since he considers that the ability to restore texts is somewhat independent of competence in a language.

The possibility of improving the reliability and validity of a cloze procedure by applying traditional item analysis and selection techniques is discussed by Brown (1988), who uses classical item analysis techniques to select the best option on the basis of item facility and discrimination indices. Brown (1993) is also concerned with the characteristics of natural cloze tests: scoring methods, length of blanks, frequency of deletions, passage readability, native and non-native performance, and test length are the variables considered. Jonz (1991) takes the cloze item types across the boundaries of the sentence. In his research it is found that intersentential ties are particularly salient in the comprehension process of nonnative speakers and consequently fixed-ratio cloze tests are significantly sensitive to textual variations and continuities at levels well beyond local phrase structure.

b) Versions of clozes

1. C-test

In 1981, Christine Klein-Braley and Ulrich Raatz introduced a new deletion technique, "the rule of 2," which was believed to remedy most of the shortcomings of the classical cloze procedure. According to "the rule of 2," the second half of every second word should be deleted in a test, starting and ending with an intact sentence. Rather than language, Klein-Braley and Raatz (1984) are concerned with the testing technique itself. The former discusses the Classical Latent Additive Test Model (CLA Model) in which the item difficulties and subject abilities can be estimated independently of each other. It has some of the main characteristics of the Rasch Model and is presented as an alternative to the classical discrete-point item tests. In an effort to validate the C-test, Klein-Braley (1985) tries to present her C-Tests as technically superior to cloze tests.

The C-test is evaluated against four different language test among Hungarian EFL learners. Dornyei and Katona (1992) confirm that the C-test is a reliable and valid instrument, and that detailed information can be obtained about issues such as text difficulty and text appropriateness, the role of content and structure words, and the use of different scoring methods.

2. The letter-deletion procedure (LDP): a number of letters may remain undeleted at the beginning of item words; the number varies from 0 to about n/2, when *n* is the number of letters in the item word, depending on the contextuality groups of the rational deletion system, on item system and on item word length. With this technique Kokkota (1988) tries to overcome the scoring problems inherent in the cloze system and the deletion inflexibility of the C-test. His conclusion is that his letter-deletion procedure (LDP) is a more flexible and powerful means of controlling reduction of text redundancy than cloze procedure or the C-test.

c) Multiple choice

It seemed that Taylor's cloze was the answer to all the shortcomings of the multiple choice. Four decades have passed since the cloze technique came out and the multiple choice, in spite of its many detractors, is still used. Economical reasons in its administration, the ease with which it is computerized, or its advantages, especially in reading comprehension, could be the reason for the support it claims.

Sang et al. (1986) recur to the multiple choice technique to confront the unitary competence hypothesis (Oller 1976) with new evidence supporting a multidimensional model of foreign language ability. Their hypotheses were tested using confirmatory factor analysis, but the seven tests (elementary: pronouncing, spelling, lexicon; complex: grammar, reading comprehension; communicative: listening comprehension, interaction) were presented in a multiple-choice form.

Chapelle (1988) studies the relationship between field independence and language measurements to compare the different techniques. She recurs not only to cloze and dictation techniques but also to the multiple-choice language tests. Two years later, as an element of contrast, she introduced the multiple choice again. This time a comparative study was carried out between four different procedures: fixed ratio/rational, multiple choice and C-test. Bachman and Palmer (1989) in their construct validation of self-rating of communicative language ability research use a 21-item multiple-choice self-rating test. Finally, Allan (1992) recurs to this technique in his elaboration of a scale to measure test-wiseness.

Looking for alternative procedures to the multiple choice, Meara and Buxton (1987) discuss the multiple choice technique versus Yes / No questions. They present the Y/N technique as an alternative to multiple choice vocabulary tests. The results obtained suggest advantages over the more traditional multiple choice format for testing vocabulary. Jafarpur (1987) studies some of the traditional criticisms on reading tests and the ways in which an alternative approach—namely, the short-context technique—avoids this defect, though many readers saw it simply as a more contextualised multiple choice. In spite of the advantages of both the Yes / No question and the short-context technique, the student is still required to make a choice and discriminate between alternatives. Finally, in his attempt to present a valid measurement of monitored knowledge Dekeyser (1990) argues that a fill-in-the-blanks format is to be preferred over multiple choice, grammaticality judgement or error correction tasks. Again, not much change is observed in this format in relation to the multiple choice format since it still requires the testee to recover a text, though under guidance. The student does not have to compose or construct an answer either.

9. WHAT WE HAVE FOUND

At this juncture it might be appropriate to comment on the journal's timeliness and the coherence of its editorial policy in the ten years that followed. Although it is probable that most of the topics *LT* has dealt with could have been read or published in a variety of journals, to have a specific forum owing to an ever-growing interest in assessment was justification enough for the launching of this publication ten years ago. The fulfilling of its aims and scope justified its first decade and will no doubt justify its second, since the range of topics discussed has included theoretical issues and empirical research in the domain of the assessment of language ability. It has published articles on:

—research into different batteries of standard tests: EPTB, ELTS or TOEFL
—research into methods of testing: introspection and computer-assisted self-assessment
—research into the different testing techniques: multiple choice, translation, clozes and C-tests
—research into test analysis: the CTT and the IRT
—and attempts to test comunicative competence.[5]

Various questions could be addressed in a discussion of the relevance of this journal and all sorts of answers could be expected. Its detractors might wonder about its specific contribution to the scientific community while its supporters will find it relevant and indispensable. What nobody will dispute is that researchers in this field no longer need to scan the summaries of all the possible journals for articles of the type published in *Language Testing.* The topics discussed in the issues published over these ten years amount to a real state of the art in language testing; *LT* goes beyond the scope of other publications in the same field such as the *Journal of Educational Measurement,* whose concern is to promote greater understanding and improved use of measurement techniques in education rather than in the language domain. Specific though *LT* is, however, it ranges over the following topics: the main features of testing: validity and reliability, item response theory, authenticity—of tasks, texts, content and setting—strengths and weaknesses of the different testing techniques, nonlinguistic issues on testing, cultural background and test tasks strategies, and self-assessment of language proficiency. All of them are topics which we have to cope with whether we

are to assess English for General Purposes or English for Specific Purposes, the latter having been largely neglected heretofore in most of our syllabi

10. WHAT WE MISS

—It would be helpful to find something similar to the instructional modules on issues in educational measurement—ITEMS—published by the National Council of Measurement in Education of USA (NCME). Presumably, quite a large number of people concerned with language testing have an Arts background, and although there is quite a lot of literature on the skills in the use of statistics needed for language studies—Butler (1985), Wood et al. (1986), Hatch and Lazaraton (1991), and Weir and Roberts (1994), among others—many readers of *LT* would appreciate some clear examples of how statistics may be used, a matter that is beyond the average reader. Some sort of self-test of the ITEMS type would be welcome especially in those articles that take for granted an advanced knowledge of statistics.

—There are also assumptions on the researcher's side which can mean gaps of information for the reader, since the topic or experiment may fall quite outside his/her field of research, as happens in the IRT contributions, where the researcher thinks that the reader is fundamentally familiar with the models of the latent trait theory, the item characteristic curve, or with the agreement coefficient, kappa coefficient, phi (lambda) dependability index or the short-out method phi coefficient, and for this reason omits material relevant to the final interpretation.

—Most of the empirical research articles offer explanatory appendixes, though, if not an expert, the reader may miss a fuller explanation. This is the case with some scoring processes, especially where the C-test technique is applied.

—We also miss more empirical work on Second Language Acquisition, which paradoxically is a point of reference in most theoretical research concerned with general issues rather than the daily needs that come up in our testing activities. And it must be borne in mind when we take the typical standard batteries of tests as landmarks of testing that neither the cultural background nor the language of "authentic tests" are a true representation of real life cultural backgrounds and language, as claimed by Spolski: “any lan-

guage test is by its very nature inauthentic for the test taker is being asked not to answer a question giving information but to display knowledge and skill" (1985: 33). They may, however, be taken as data for research, since our classrooms are not an appropriate setting for such research because of the number of test takers required. Given its increasing relevance, little empirical research into ESP has been done over this decade.

These and other shortcomings will probably be overcome in the issues of the coming decade. We see grounds for optimism in the editorial board policy, which after some hesitation about whether or not to change the name of the journal has opted for an expansion of content rather than a change of title. Moreover, the average reader's literacy in statistics may be higher in the near future if applied statistics becomes a compulsory subject in our curricula. For the time being, however, we still need some guidelines on how to test our hypotheses and how to interpret the data we get, despite all the help available to most of us from the computer.

In our opinion, then, *LT* has made a notable contribution over the last 10 years to the debate about how far language testing has gone toward understanding the abilities that teachers and institutions intend to measure. It is high time we, as teachers/testers of Second Language Acquisition or of Language for Specific Purposes, took advantage of these ten years of language testing research and that lamentations like the following one by Alderson (1988: 87) were progressively outdated: *"It is rather sobering and perhaps depressing to note the minimal attention paid to testing. . . ."* a

NOTES

1. Differences between norm-referenced test (NRT) and criterion-referenced test (CRT) are mainly based on type of measurement and type of interpretation, other features such as score distribution, purpose of testing and knowledge of questions are considered. (from Brown 1990: 79)

2. The interest rests upon the individual items of a test rather than upon some aggregate of the item responses such as a test score. A reasonable assumption is that each examinee responding to a test item possesses some amount of the underlying ability tested and that at each

ability level there will be a certain probability that an examinee with that ability will give a correct answer to the item. The concept of test reliability is replaced in item response theory by the vastly superior concept of test information.

3. The Partial Credit model is an extension of the simple Rasch dichotomous model (Rasch, 1960, 1980) that allows for the scoring of items in any number of ordered categories. The dichotomously scored test items give way to a rating of 0, 1, 2... according to its degree of increasing acceptability and appropriateness.

4. Equating: a technical term in testing literature, which involves administering a small set of items with an older form as well as the new one in order to identify comparable score levels.

5. While most areas of grammatical competence can and have already been translated into operational, dichotomous testing items, the complexity in translating components of communicative competence into testing items still persists: See Olshtain and Blum-Kulka (1985), Bachman and Palmer (1989), and Swain (1993) among others.

ABBREVIATIONS

ANOVA	Analysis of variance
CFA	Confirmatory factor analysis
CLA	Classical Latent Additive Test Model
CRM	Criterion-referenced measurement
CTT	Classical Test Theory
EFL	English Foreign Language
ESP	English for Specific Purposes
ELBA	English Language Battery
ELTS	English Language Testing Service
EPTB	English Proficiency Test Battery
IRT	Item Response Theory
LTD	Letter Deletion Model
MTMM	Multitrait Multimethod
NCME	National Council of Measurement in Education
NRM	Norm -referenced measurement
LT	Language Testing Journal
MC	Multiple Choice
PCF	Person characteristic function

PIRC	Proportional Item Response Curve
TOEFL	Testing of English as a Foreign Language
TSE	Test of Spoken English
TWE	Test of Written English

WORKS CITED

ADAMS, R. J. et al. 1987. "A Latent Trait Method for Measuring a Dimension in Second Language Proficiency." *Language Testing* (*LT*) 4: 9-27.

ALDERSON, J. C. 1988. "Testing and its Administration in ESP." In *ESP in the Classroom: Practice and Evaluation.* (ELT Document 128). Ed. D. Chamberlain and R. J. Baumgardner. Modern English Publications / The British Council. 87-97.

ALLAN, A. 1992. "Development and Validation of a Scale to Measure Test-Wiseness in EFL/ESL Reading Test Takers." *LT* 9: 101-122.

AMER, A. A. 1993. "Teaching EFL Students to Use a Test-Taking Strategy." *LT* 10: 71-78.

BACHMAN, L. F., et al. 1988. "Task and Ability Analysis as a Basis for Examining Content and Construct Comparability in Two EFL Proficiency Test Batteries." *LT* 5: 128-159.

BACHMAN, L. F., and A. S. PALMER. 1989. "The Construct Validation of Self-Rating of Communicative Language Ability." *LT* 6: 14-29.

BOLDT, R. F. 1989. "Latent Structure Analysis of the Test of English as a Foreign Language." *LT* 6: 123-151.

- - -. 1992. "Crossvalidation of Item Response Curve Models Using TOEFL Data." *LT* 9: 79-100.

BRADSHAW, J. 1990. "Test-Takers' Reactions to a Placement Test." *LT* 7: 13-30.

BROWN, J. D. 1988. "Tailored Cloze Improved with Classical Item Analysis Techniques." *LT*, *LT* 5, 19-48.

- - -. 1990. "Short-cut Estimators of Criterion-referenced Test Consistency." *LT* 7: 77-97.

- - -. 1993. "What Are the Characteristics of Natural Cloze Tests?" *LT* 10: 93-116.

BUTLER, C. 1985. *Statistics in Linguistics.* Oxford: Blackwell.

CARMAN, R. A., and W. R. ADAMS. 1972. *Study Skills: A Student's Guide for Survival.* New York: Wiley.

CARROLL, J. 1986. "LT+ 25, and Beyond? Comments." *LT* 3: 123-130.

CHAPELLE, C. A. 1988. "Field Independence: A Source of Language Test Variance?" *LT* 5: 62-82.

CHIHARA, T. et al. 1989. "Background and Culture as Factors in EFL Reading Comprehension." *LT* 6: 143-151.

CHOI, I., and L. F. BACHMAN. 1992. "An Investigation into the Adequacy of Three IRT Models for Data from two EFL Reading Tests." *LT* 9: 51-78.

DAVIDSON, F., and G. HENNING. 1985. "A Self-Rating Scale of E. Difficulty: Rasch Scalar Analysis of Items and Rating Categories." *LT* 2: 164-179.

DAVIES, A. 1984. "Validating Three Tests of English Language Proficiency." *LT* 1,: 50-69.

DE JON, H. A. L., and C. A. W. GLASS. 1987. "Validation of Listening Comprehension Tests Using Item Response Theory." *LT* 4: 170-194.

DE MAURO, G. 1992. "Examination of the Relationships among TSE, TWE and TOEFL Scores." *LT* 9: 149-162.

DEKEYSER, R. 1990. "Towards a Valid Measurement of Monitored Knowledge." *LT* 7: 147-157.

DORNYEI, Z., and L. KATONA. 1992. "Validation of the C-test amongst Hungarian EFL Learners." *LT* 9: 187-206.

FREEDLE, R., and I. KOSTIS. 1993. "The Prediction of TOEFL Reading item Difficulty: Implications for Construct Validity." *LT* 10: 133-170.

HALE, G. A. 1988. "Student Major Field and Text Content: Interactive Effects on Reading Comprehension in the Test of English as a Foreign Language." *LT* 5: 49-61.

HALE, G. A., et al. 1989. "The Relation of Multiple-choice Cloze Items to the Test of English as a Foreign Language." *LT* 6: 47-76.

HATCH, E., and A. LAZARATON. 1991. *The Research Manual: Design and Statistics for Applied Linguistics.* New York: Newbury House.

HAUGHTON, D., and L. DICKINSON. 1988. "Collaborative Assessment by Marking Candidates in a Tutor Based System." *LT* 5: 233-249.

HEILENMAN, L. K. 1990. "Self-Assessment of Second Language Ability: The Role of Response Effects." *LT* 7: 174-201:

HENNING, G. 1984. "Advantages of Latent Trait Measurement in Language Testing." *LT* 1: 123-133.

- - -. 1988. "The Influence of Test and Sample Dimensionality on Latent Trait Person ability Item Difficulty Calibrations." *LT* 5: 83-99.

- - -. 1989. "Meanings and Implications of the Principle of Local Independence." *LT* 6: 95-118.

HENNING, G. et al. 1985. "Item Response Theory and the Assumption of Unidimensionality for Language Tests." *LT* 2: 141-154.

HUDSON, T. 1993. "Surrogate Indices for Item Information Functions in Criterion-Referenced Language Testing." *LT* 10: 171-192.

HUDSON, T., and B. LYNCH. 1984. "A Criterion-referenced Measurement Approach to ESL Achievement Testing. *LT* 1: 171-201.

JAFARPUR, A. 1987. "The Short-context Technique: An Alternative for Testing Reading." *LT* 4: 195-224.

JANSEEN, V. D. 1989. "The Development of a Test of Dutch as a Second Language: The Validity of Self- Assessment by Inexperienced Subjects." *LT* 6: 30-46.

JONZ, J. 1991. "Cloze Item Types and Second Language Comprehension." *LT* 8: 1-22.

KLEIN-BRALEY, C. 1985. "A Cloze-Up on the C-Test: A Study in the Construct Validation of Authentic Tests." *LT* 2: 76-104.

KLEIN-BRALEY, C., and U. RAATZ. 1984. "A Survey of Research on the C-Test." *LT* 1: 134-46.

KRZANOWSKI, W. J., and A. WOODS. 1984. "Statistical Aspects of Reliability in Language Testing." *LT* 1: 1-20.

KOKKOTA, V. 1988. "Letter Deletion Procedure: a Flexible Way of Reducing Text Redundancy." *LT* 5: 111-126.

LADO, R. 1986. "Analysis of Native Speaker Performance on a Cloze Test." *LT* 3: 130-146.

MCNAMARA, T. F. 1990. "Item Response Theory and the Validation of an ESP Test for Health Professionals." *LT* 7: 52-76.

- - -. 1991. "Test Dimensionality: IRT Analysis of ESP Listening Test." *LT* 8: 139-159.

MEARA, P., and B. BUXTON. 1987. "An Alternative to Multiple Choice Vocabulary Tests." *LT* 4: 142-154.

OLLER, J. W., and C. A. CONRAD. 1975. "The Cloze Technique and ESL Proficiency." *Language Learning* 21: 183-95.

OLSHTAIN, E., and S. BLUM-KULKA. 1985. "Crosscultural Pragmatics and the Testing of Communicative Competence." *LT* 2: 16-30.

OSCARSSON, M. 1989. "Self-assessment of Language Proficiency: Rationale and Application." *LT* 6: 1-13.

PERKINS, K., and L. D. MILLER. 1984. "Comparative Analyses on Item Responses." *LT* 1: 21-32.

POLLIT, A., and C. HUTCHINSON. 1987. "Calibrating Graded Assessments: Rasch Partial Credit Analysis of Performance in Writing." *LT* 4: 72-98.

POWERS, D. E. "Academic Demands Related to Listening Skills." *LT* 3: 1-39.

RASCH, G. 1980. *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago: U of Chicago P. (First pub. Copenhagen: Danish Institute for Educational Research, 1960).

RYAN, K. E. and L. F. BACHMAN 1992. "Differential Item Functioning on Two Tests of EFL Proficiency." *LT* 9: 12-29.

SANG, F., et al. 1986. “Models of Second Language Competence: A Structural Equation Approach.” *LT* 3: 54-79.
SELIGER, H. W. 1985. “Testing Authentic Language: The Problem of Meaning.” *LT* 2: 1-15.
SCOTT, M. L. 1986. “Student Affective Reactions to Oral Language Tests.” *LT* 3: 99-122.
SHOHAMY, E. and T. REVES. 1985. “Authentic Language Texts, Where from and Where to?” *LT* 2: 48-59.
SPOLSKY, B. 1985. “The Limits of Authenticity in Language Testing.” *LT* 2, 31-40.
SPOLSKY, B. 1990. “The Prehistory of TOEFL.” *LT* 7: 98-118.
SWAIN, M. 1993. “Second Language Testing and Second Language Acquisition: Is There a Conflict with Traditional Psychometrics?” *LT* 10: 193-210.
THEUNISSEN, T. J. J. M. 1987. “Text Banking and Test Design.” *LT* 4: 1-27.
TOMLINSON, B., et al. 1988. “An Algorithmic Approach to Prescriptive Assessment in English as a Second Language.” *LT* 5: 1-18.
WEIR, C., and J. ROBERTS. 1994. *Evaluation in ELT.* Oxford, Blackwell.
WOODS, A., and R. BAKER. 1985. “Item Response Theory.” *LT* 2: 119-140.
WOODS, A., et al. 1986. *Statistics in Language Studies.* Cambridge: Cambridge: Cambridge UP.
ZEIDNER, M. 1986. “Are English Lnguage Aptitudes Tests Biased Toward Culturally Different Minority Groups? Some Israeli Finding.” *LT* 3: 80-98.
Zeidner, M., and M. Bensoussan. 1988. “College Students’ Attitudes towards Written versus Oral Test of English as a Foreign Language.” *LT* 5: 100-127.

a