QUALITY AND QUANTITY IN ENGLISH LINGUISTICS RESEARCH:

Some issues

Edited by

Cesare Gagliardi

LIBRERIA DELL'UNIVERSITÀ EDITRICE
PESCARA 2002

To Ula, brave and intense

11.

RICERCHE LINGUISTICHE
Series directed by Cesare Gagliardi

INDEX

Nunan D. 1991. "Methods in second language classroom-oriented research: A critical review". *Studies in Second Language Acquisition* 13, 249-274.

Polio C. & S. Gass 1997. "Replication and reporting: A commentary". *Studies in Second Language Acquisition* 19, 499-508.

Scollon R. 1995. "From sentences to discourses, ethnography to ethnographic: Conflicting trends in TESOL research". *TESOL Quarterly* 29, 381-384.

Thomas M. 1994. "Assessment of L2 proficiency in second language acquisition research". *Language Learning* 44, 307-336.

QUANTITY AND QUALITY IN KNOWLEDGE SPACES FOR FOREIGN LANGUAGE COMPUTER-BASED TESTING

Anna Zanfei

Psychometric reliability and therefore quantitative statistical research methodology has been carried out together with qualitative approaches since the 1980s. The latter investigates a wide variety of factors in test-development – such as field dependence and discourse domain definition – as well as test-taking processes – such as background knowledge effects and any strategy used that could influence test performance.

Nevertheless, as L. F. Bachman (2000) states, it was only towards the end of the 1980s that language test designers started to take the language learner's developmental sequence in the interpretation of test scores into account. This meant a broadened view of language test constructs which led to the merging of language testing research into applied linguistics research.

As far as research methodology in language testing is concerned, techniques have become more and more sophisticated and diverse. Classical test theory and factor analysis are still in use, however, criterion-referenced measurement, generalizability-theory – known as G-theory –, item response theory and structural equation modelling are generally considered the most powerful quantitative methodologies.

Throughout Europe and especially in Italy a criterion-referenced measurement with special reference to the *Common*

European Framework (1996) is widely used. Test results are thus interpreted with reference to the six levels described in the *Framework* even if norm-referenced measurement procedures are used complementarily with criterion-reference ones.

Bachman (2000) explains how the G-theory is considered as an extension of classical test theory that enables the language testing researchers to estimate multiple sources of measurement error and dependability concerning a criterion. Item Response Theory is currently used for large-scale standardized tests and for computerized adaptive tests (CATs). The IRT model shows the statistical properties of items as well as the test takers abilities estimation. This statistical measure is thus sample-free and test-form-independent. A limitation of IRT is that it is only applicable to dichotomic scoring: an item is scored as right or wrong. This means that task-responses expected to be partially right – from a semantical or syntactical or pragmatical point of view – cannot be employed in the assessment procedure.

Regression and factor analysis have been extended resulting in Structural Equation Modelling (SEM), which provides statistical evidence for investigating the factor structure of the measures used and how those factors are related. Dependent and independent sets of factors can show directional relationships that are revealed through this analysis. Test takers background, their personal characteristics, the strategies they adopt, their language proficiency and cognitive abilities, have been investigated in terms of their influence on test performance using Structural Equation Modelling. Test researchers use other quantitative tools such as multidimensional scaling. As Bachman (2000) states, latent trait approaches are leading to a single analytic paradigm that will bring all the quantitative methods together.

In language testing research, qualitative methods have been used to investigate test taker characteristics, their processes and strategies in answering test-tasks and the discourse they create in productive tasks. Data are collected through expert judgements, verbal reports, systematic observations, questionnaires, interviews and, in addition, discourse analysis techniques. Self-assessment and direct observation have become part of the testing procedure that includes a wide variety of productive and receptive tasks. The variability of social properties in speech events and the speaker's choices used for achieving communicative goals, as well as the different types of data collections procedures and methods, are now becoming the focus of pragmatic competence in cross-cultural communication research. Corpora of writing tasks previously marked by human raters have proved to be very useful in identifying the main features of this genre of discourse. English for Specific Purposes testing procedures are based on the research results about the effects of topic knowledge on test performances. The ESP tests require a good knowledge of the subject matter and a specific training in the mastering of the language use relevant to each specific domain. The *test construct* is thus narrowed to a well-defined domain and therefore made clearer.

A framework for L2 vocabulary assessment that turned out to be a good instrument for evaluating the validity of existing tests is now available. This has led to an application of interactionalist principles in vocabulary test constructs and hence in the interpretation of scores. Read and Chapelle (1999) state that the distinction between different theoretical perspectives on vocabulary becomes evident only through the examination of data produced by distinct forms of assessment. The mental lexicon has been investigated in association with the instrument

of the *c-test*, which focus on a number of linguistic features in texts. On the contrary, the use of decontextualized items to test knowledge of content words is linked to the treatment of vocabulary as a separate construct. Another way to investigate vocabulary is through the analysis of lexical density statistics and of the occurrence of lexicalized phrases found in written extended performances. The psychometric practice validation of these testing procedures is still found by means of a correlation with a criterion measure. Inferences, made from the proportion of correct responses to items, are made at subtest level when the vocabulary measure is embedded in a larger test and at whole test level when the learner's writing performance is concerned, as is also a general measure of the learner's vocabulary knowledge.

A positive development in testing practicality is due to the environment offered by computer technology that permits the use of multimedia sources delivered as *input* and a straightforward statistical computation of data. Data are statistically processed in real time; listening (video and audio) texts are easily delivered; photos and coloured drawings are true copies of their original source. An original text can be scanned and reproduced in a digital form so authenticity is substantially respected in that way. In addition students can write short answer texts that are automatically matched with all the possible anticipated responses and automatically scored as correct or partially correct. Finally while automatic "essay graders" are precious tools they are in use only through a limited testing agencies.

The construct of computer based language tests should take into consideration its new task context in relation to the real life language use context.

CBT architecture. H. Braun (1999) uses the metaphor of architectural design in order to explain the test design process. As a matter of fact, test designers employ criteria that are similar to those employed by architects and in CBT this dimension is even more emphasised. For example, a CBT produced with an authoring system absorbs its architectural potentialities, hence the importance of open scripting architecture software programs. Through OSA authoring systems, the layout of input materials is easily customized, objects (digitalized videos, audios, images) and programs are easily embedded into the CBT, and user-customized testing procedures are based on user actions – e.g. responses.

The language use domains taken into consideration are the starting point for any test design. This immediately suggests similarities with architectural design procedures. The natural landscape is for the architect what a language domain is for the test designer. The elements of test design are items which, in CBT environment for example, have the property of employing multimedia sources. The features of setting, developement and delivering are constraints which set the rules for the actual construction of tests. Constraints have to do with measurement – distribution of difficulty, reliability, comparability, generalizability in the testing environment; modes – computer-delivering and automatic scoring in CBT case; procedures – test-method facets; and probability psychometric tools which add consistence to the construction. All the above mentioned constraint components play an important role in CBT. Obviously, the design purpose influences the salience of criteria used. Measurement, as far as language testing is concerned, is the process of quantifying test-taker performances. Tests can be used as instruments of measurement that lead an expert to make

inferences about test-taker's language abilities when they are calibrated on a scale. Finally, if the test is referred to a criterion, the purpose would be to state whether or not specified skills or abilities have been mastered. Factors that affect language test scores have been illustrated by Bachman (1991) as: language ability under examination, test method facets, candidates' personal attributes and random factors. These factors and hence their treatment for reliability and validity do not change with computer delivered tests. One interesting aspect of computer delivered tests is the great amount of data, useful for quantitative as well as qualitative investigation, that can be gathered in a non-pervasive manner. Time taken, outcomes chosen or written, personal data and other unscored items can be tracked by directly or indirectly queries. In addition, every time a test is taken it can be recorded as part of an individual learning curriculum.

CBT is a productive environment for research because of the ease of gathering a lot of information, of using multimedia sources and of applying new psychometric approaches involving complex statistical procedures that need to be computed in real time. One of these approaches is knowledge spaces theory which has been applied to the interaction between the knowledge domain under consideration and the language learner developmental domain.

Language learner dynamic domain has also been investigated by Buck and Tatsuoka who applied a rule-space methodology to listening test data (Buck & Tatsuoka 1997).

> Rule space methods, when successfully applied, allow cognitively based interpretations of test performance that meaningfully differentiate among individuals at different score levels and even among individuals at similar score levels but

> with qualitatively different response patterns. (Braun 1999:267)

The focus is here on the actual state of language knowledge and its prerequisites. Two students with the same test final score may yield different response patterns showing that different prerequisites are missing. This can be of little importance for the time being, but it can be prejudicial to further development of the individual learning process.

Knowledge space theory has been applied to procedures for the adaptive assessment of knowledge, by using a large set of items from a specified field of interest and prerequisite relationships between these items. These relationships form a networked structure that links the items and this structure turns out to be useful for adapting the item presentation to students according to her/his true ability. The resulting computer adaptive test is thus tailored to users in the sense that items which are too easy or too difficult are avoided in the actual test procedure.

The latter adaptive procedure differs from that derived from the application of IRT in CATs; firstly, because it is drawn from a computerized procedure for querying experts on prerequisite relationships and, secondly, because it looks for a local dependence among items where IRT CATs, on the contrary, rely on item-independence. The need for a computerized query is due to the fact that the possible prerequisite relationships among items pertaining to a domain grow exponentially with the number of items involved.

An example can be a very small set of items like the following:

1. is being able to choose the right synonym out of three given, of a word underlined in a text.
2. is being able to choose the best statements given (4 out of 16) which paraphrase the detailed information in a written text.
3. is being able to choose the best statements given (4 out of 16) which give the gist of a listening text.
4. is being able to reconstruct a text with the help of grammatical and lexical knowledge and the comprehension of the text discourse by filling in the gaps in a cloze text choosing words and expression from a list containing several distractors.

Another easier way is to use a simple database composed of discrete grammatical multiple choice items in which a prerequisite relationship structure between items is feasible.

A different perspective is the one developed throughout the *European framework*:

> For example on a self-assessment checklist, the descriptor *Can ask for and provide personal information* on the sub-scale *Exchanging Information* might be divided into the following implicit constituent parts:
>
> 1. I can introduce myself;
> 2. I can say where I live;
> 3. I can say my address in English,
> 4. I can say how old I am, etc. and I can ask someone what their name is;
> 5. I can ask someone where they live,
> 6. I can ask someone how old they are. (E. C. F. of Ref. 8.4)

An expert – e.g. a teacher – can judge which of the above tasks are more or less difficult taking into account the syntactical prerequisites involved in each one and comparing this with all the others. A rank scale order can be thus created. At this point it is easy to infer that a learner who cannot – when speaking in the target foreign language – introduce himself/herself is even more likely to have difficulty in asking someone else their name or where they live or in uttering her/his own address.

The *European framework* describes six levels of competence in a scheme that starts from an initial division into three broad levels that branch into lower and finer levels taking the shape of trees that interrelate with one another.

A basic user level has two lower level branches (a1) or *Breakthrough* and (a2) or *Waystage* which divide up into lower levels (a1.1, a1.2) and (a2.1, a2.2, a2.1.1, a2.1.2). All these are states that can be described in terms of "can do" statement concerning reading, listening, writing and speaking with reference to a domain and a category of situations.

This is a way of thinking about language as linked to concrete situations, context of reality and concrete or abstract contents. Prerequisites concerning language tasks typical of activities related to specific situations are derived from the analysis of task characteristics as Bachman and Palmer state:

> ... language use is embedded in particular situations which may vary in numerous ways... it is, however, possible to identify certain distinguishing characteristics of language tasks and to use these characteristics to describe a language use domain. (Bachman & Palmer 1996:44)

The degree of correspondence between target language use tasks and test tasks is then defined by Bachman and Palmer by setting out test tasks in terms of their characteristics:

> i.e. multiple choice [*items vary in terms of*] syntactic complexity, length, level of vocabulary, topical content, type of response required. (Bachman & Palmer 1996:46)

The framework defined by Bachman and Palmer should be of help in discovering the minimum prerequisites necessary to solve a test task and eventually write prerequisite-items classified as enabling skills and basic knowledge.

The prerequisite relationships can be represented as an "*and/or* graph" that defines all the possible paths that lead to the solution of a complex "can do" statement.

From an incorrect answer to one or more nodal "enabling items" it can be inferred that it is very likely that a student would not know how to solve tasks which include such items. Conversely, we can easily visualize in graphic form those skills which are missing and are also preventing a student from successfully mastering any complex task. Furthermore, we can investigate the actual states of knowledge that compose the space of a target language domain.

The language domain space is validated through the comparison between an empirical space built on the basis of test data and the model derived from the expert's query.

How knowledge space query works. The first phase of the process that lead to the construction of a knowledge space consists in creating a list of tasks – or questions or items or problems – that cover a specific language domain. This list is

simply called a domain and it should contain complex tasks as well as their prerequisites. The items are then ranked and at this point a querying procedure can take place. The standard question form displayed by a "knowledge space query computer program" is :

> "Imagine that a student does not master the items $a_1,....a_k$. IS IT PRACTICALLY CERTAIN that this student will not master item q?"
>
> $a_1,....a_k$ are called the premise and item q is the consequence. (Doignon & Falmagne: 1999)

Studies that validate experts' knowledge spaces on first language reading and writing abilities (Albert, Dowling & Riesenhuber: 1994) and that validate empirical knowledge space for foreign language have been carried out in Austrian and German Universities (Duntsch & Gediga: 1997).

The purpose of querying experts is to identify a space within which there is agreement, usually among three or more experts, and validate this model by comparing it with data provided by a test administered to a representative sample.

Once the most likely answer patterns are found, an algorithm can be constructed. The implementation of this algorithm in a CAT enables the program to eliminate automatically all those patterns that contain an item which has been answered incorrectly. Any question which follows the first is thus chosen at random among those that mostly occur in the set of patterns that have not been excluded. Every time an answer is given the set of patterns available is divided into two: one is excluded and the other is used to continue the test. This operation shortens the time required to take the test by reducing the number of questions needed to determine the knowledge state of a test-taker. In the

end, when the testing procedure finds a pattern that corresponds to the real knowledge state of the test taker, this state is tested through further questions chosen by a rule of neighbourhood.

It can be concluded that, as J. Lukas and D. Albert (1999:3) point out, the knowledge space theory is submitted to interpretation within various knowledge domains and it is applicable to diagnostic as well as tutoring procedures. The most difficult step in its application in diagnostic procedures concerning English as a foreign language, seems to be the construction of a suitable list of tasks which show any prerequisite relationships and at the same time could characterize a specific language domain.

Conclusion: Quality and quantity are two complementary faces of probabilistic testing instruments by which researchers aim to discover the knowledge states of learners. Quantitative research methods should assure the consistence of the instrument by which a test-taker has been evaluated and, therefore, its reliability. The application of qualitative methods is also unavoidable in a validation procedure of such instruments. Furthermore, they can add new hypothesis about what is not immediately interpretable through numbers but is still present in the data. Besides the above mentioned research procedures, computer based testing – CBT – also needs to be comparable with other instruments.

Seen as a practical tool, CBT has never been considered a new research environment for language testing, though Chaloub-Deville (2000) and Braun (1999) give us an idea of the impact that technology is having on elicitation procedures. The mode of delivering CBTs (through local and Internet network), the convergence of different media-sources, and the gathering of

variegated data give the impression of a broadened test-context. The complementary use of corpora and word-processors have made accurate automatic analysis of limited and extended writing materials feasible. Finally since e-mailing, web-chatting, web-reading, digital audio-visual television programmes and word-processors are everyday channels of language use, these issues should be in some way taken into account in language testing.

Two kinds of language testing procedures are preferable for the simplicity of their constructs. The first is ESP testing which relies on a hypothesis that human beings are rather specialized in their language use as long as their subject matter knowledge is specialized. The second one is knowledge space theory which draws on the non-numerical treatment of the component parts of a test and of answering patterns. One application of this theory is the previously described querying procedure of experts by which a judgement should be drawn from a negative question. Unfortunately this procedure is misleading since it starts from a negative premise which can easily produce false reasoning consequences. Furthermore the prerequisites and the level of difficulty are used as synonyms throughout the knowledge space literature but the relationship between them is unclear. Concerning foreign language tests, items that have high facility values are not necessarily prerequisites of any difficult task. The only time this event might occur is when two items taken into account are locally dependent. However this has not always been made clear. Apart from the above considerations it must be added that the knowledge space theory is a new way of thinking about a computer adaptive test and of computer based tests in general. Further qualitative and quantitative research will be able to highlight the advantages and disadvantages of its application to language testing.

Bibliographical references

Albert D. & C. Dowling. 1994. "Validating expert knowledge on reading and writing abilities of children". In Pawlik, pp. 23-48.

Bachman L. F. 1991. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bachman L.F. 2000. "Modern language testing at the turn of the century: Assuring that what we count counts". *Language Testing* 17/1, 1-42.

Bachman L.F. & A. S. Palmer. 1996. *Language Testing in Practice*. Oxford: Oxford University Press.

Braun H. 2000. "A Post modern view of the problem of assessment". In Kunnan, pp. 263-272.

Buck G. & K. Tatsuoka. 1998. "The sub skills reading: Rule Space analysis of a multiple choice test of second language reading comprehension". *Language Testing 15 Research Colloquium.*

Chaloub-Deville M. (ed.). *Issues in Computer-Adaptive Testing of Reading Proficiency*. Cambridge: Cambridge University Press.

Cook G. 1989. *Discourse*. Oxford: Oxford University Press.

Doignon J. P. & J.C. Falmagne. 1999. *Knowledge Spaces*. Berlin: Springer-Verlag.

Dowling C. E. & K. Rainer. 1995. "Prerequisite relationships for the adaptive assessment of knowledge". In Greer, pp. 43-50.

Düntsch I. & G. Gediga 1998. "Knowledge structures and their application in CALL systems". In Jager, Nerbonne, van Essen, pp.177-186.

Greer J. (ed.). *Artificial Intelligence in Education*. Washington, D.C.: Association for the Advancement of Computing in Education.

Jager S., J. Nerbonne & A. van Essen (eds.). *Language Teaching and Language Technology*. Lisse: Sweets & Zeitlinger.

Koppen M. & J.P. Doignon. 1990. "How to build a knowledge space by querying an expert". *Journal of Mathematical Psychology* 34, 40-49.

Kunnan A. J. (ed.). *Fairness and Validation in Language Assessment*. Cambridge: Cambridge University Press.

Modern Languages: Learning, Teaching, Assessment. A Common European Framework of reference.
Strasbourg1996: http://culture.coe.fr/lang/eng/eedu2.4.html

Lukas J. & D. Albert. 1999. "Knowledge structures: What they are and how they can be used in cognitive psychology, test theory, and design of learning environments". In Luckas & Albert, pp. 3-12.

Luckas J. & D. Albert (eds.). *Knowledge Spaces: Theories, Empirical Research, and Applications*. London: Laurence Erlbaum Associates Publisher.

Pawlik K. (ed.). *Abstracts of the 99.th Congress of the Deutsche Gesellschaft für Psychologie*. Hamburg: Psychologisches Institut I der Universität Hamburg.

Read J. & C. A. Chapelle. 2001. "A framework for second language vocabulary assessment". *Language Testing* 18/1, 1-32.

Appendix: The interface of a query program

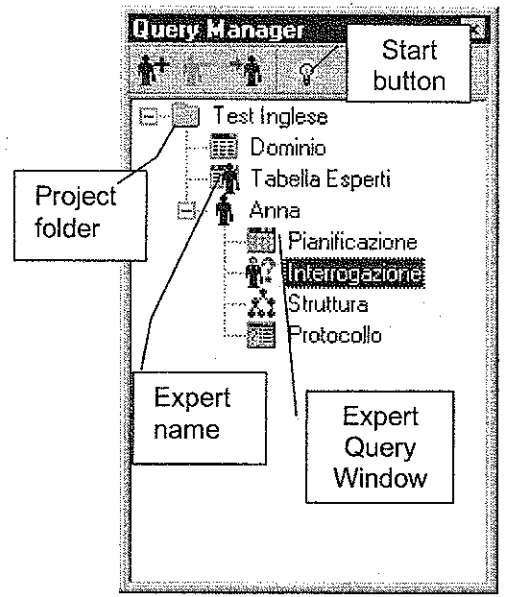

Figure **1**. Query Manager.

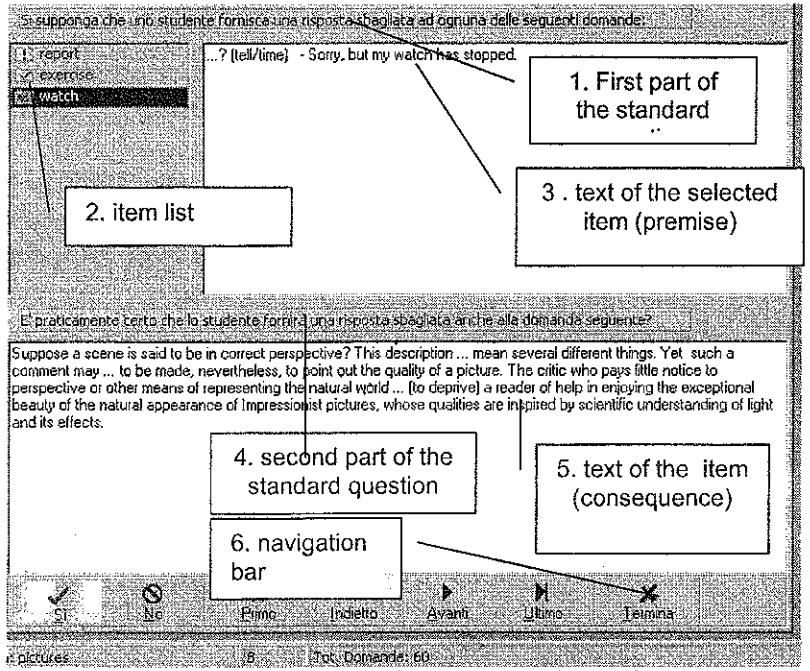

Figure 2. A query session (program XQuery)

REACHING AUTONOMY IN THE ANALYSIS OF ECONOMIC TEXTS: A HELPING HAND FROM COMPUTER SOFTWARE

Roberta Facchinetti

The task of the language teacher is to provide a context in which the learner can develop strategies for discovery - strategies through which he can learn how to learn.
(Johns 1991: 1)

1. Introduction

In the book *Self Access Systems for Language Learning*, Little posits that:

> it is an obvious but not always sufficiently emphasized fact that all learning involves change. We can conceive of this change as the addition of knowledge or the acquisition of skill, or both, but its outcome, on any normal understanding of what learning is, must be an enlargement of the learner's capacity to behave. (Little 1993: 1-2)

In the field of language acquisition, the words "enlargement of the learner's capacity to behave" can be paraphrased easily by "learning to be autonomous", so as to employ what we acquire with adequate competence in a variety of specific activities and contexts.