misuse of conjunctions, and restricted use of lexical cohesion.

## Language testing

**01–268 Alderson, J. Charles** (Lancaster U., UK (formerly on secondment to the British Council, Hungary); *Email*: c.alderson@lancaster.ac.uk), **Percsich, Richard and Szabo, Gabor.** Sequencing as an item type. *Language Testing* (London, UK), **17**, 4 (2000), 423–47.

A text's coherence clearly depends upon the way ideas are related within that text, both in terms of their logical relations, as well as the cohesive devices that show, or create, the links between ideas, across paragraphs and sentences. Thus, it would appear that part of the ability of a competent reader is to recognise the appropriate order of ideas in text, to identify cohesion and coherence in text in order to relate the ideas to each other, and to understand authorial intention with respect to the sequence of ideas. It follows from this that a potentially useful test method which might tap such abilities is to require candidates to inspect text in which the elements are out of order, and to reconstruct the original order. This, it might be supposed, would require candidates to detect the relationship among ideas, to identify cohesive devices and their interrelationships. Such test methods are, indeed, increasingly common in so-called reading tests. However, the present authors know of no reports of research into, or even descriptions of the use of, this promising task type. In this article they report on potential problems in scoring responses to sequencing tests, the development of a computer program to overcome these difficulties, and an exploration of the value of various scoring procedures.

**01–269 Al-Hazemi, Hassan.** Listening to the Yes/No vocabulary test and its impact on the recognition of words as real or non-real: a case study of Arab learners of English. *IRAL* (Berlin, Germany), **38**, 1 (2000), 89–94.

The Yes/No vocabulary test has been widely used in the past few years to estimate the vocabulary size of second language learners. Very few attempts have been made, however, to use the test with Arab learners of English. The present author has in the past used such a test and shown that native Arabic speakers appear to have an unusually low vocabulary proficiency. There may, however, be other factors contributing to this low level of vocabulary. In this paper the effect of listening to the Yes/No vocabulary test on the identification of words as real or non-real is examined. It was predicted that listening to the words read by the examiner or played on a tape recorder might produce better scores than reading them only on the test sheet. The paper and pencil version of the Yes/No vocabulary test was administered to 55 senior high school Saudi students. The results obtained were not as predicted. The findings showed no major difference in the overall scores, particularly in the Hits (the yes responses to real words) and False Alarms (the yes responses to imaginary words).

**01–270 Brohy, C.** (U. of Fribourg, Switzerland) **and Pannatier, M.**. L'évaluation dans l'enseignement bilingue: la quadrature du cercle? [Assessment in bilingual education: the squaring of the circle?] *Babylonia* (Comano, Switzerland), **1** (2000), 33–35.

Correction and assessment in a bilingual immersion situation in primary and secondary schools are discussed here. Aspects of research and practice at three levels, micro, 'mezzo' and macro, are considered. In some countries (such as Canada) a slight shift from 'focus on meaning' to 'focus on form' means that teachers lack an inventory of good correction practice. The micro-level relates to correction in the classroom, where students in an immersion situation tend to prefer not to be interrupted for corrective purposes. 'Mezzo'-level is concerned with the role of correction and assessment in tests and examinations, where there is the problem of what is being tested, the learner's use of the second language or knowledge of the subject. (There is no generally shared practice, though usually it is the subject that is assessed rather than the language; however, institutions have difficulty persuading students that this is the case.) Activities at the macro-level involve passing on an assessment to others (e.g., in reports to parents). The authors conclude that serious reflection is required on assessment in bilingual education, which could have beneficial consequences for assessment in general.

**01–271 Daniëls, John.** Kan een computer samenvatten? Een nieuw type examen? [Can a computer summarise? A new type of examination?] *Levende Talen Magazine* (Amsterdam, The Netherlands), **7** (2000), 10–12.

This article draws on one in *Computer Totaal* by Kees Vuik, who tested the expensive summary-generating program *Sinope*, by *Comsis*. The producers claim that the program can analyse texts in 'syntactically sound' parts, build a semantic tree and then prune it in a sophisticated way, leaving the most important components. However, Vuik found the program failed to deliver satisfactorily, each successive pruned version of the text losing the overall sense of the original; and he suggested an alternative (and cheaper) way of constructing one's own program, which is elaborated in the original article. He adhered to the priniciple, moreover, that the work of the computer should always be followed by human input, the combination of the two generating good summaries. Building on these ideas, the present

author suggests constructing a new type of school examination for Dutch and foreign languages. Asking a computer-student combination to summarise texts seems an excellent way to replace the present form of questioning, where what is important in the text has been decided in advance. It seems much more instructive to ask the students to question the text themselves and then to edit a computer-made summary than to focus on parts of the text in order to answer questions posed by others.

**01–272 Gohard-Radenkovie, Aline** (Fribourg U., Switzerland). Evaluer les compétences socioculturelles: le cas de l'étudiant universitaire en situation de mobilité. [Evaluating the sociocultural competencies of students involved in higher education exchange programmes.] *Babylonia* (Comano, Switzerland), **1** (2000), 41–46.

This article looks at modes of evaluation of cultural competencies acquired by students involved in higher education exchange programmes. It begins by considering how to conceptualise cultural competence and the types of knowledge and skills which need to be co-constructed with the learner during the overseas exchange visit. Two approaches stand out: on the one hand there is *declarative* evaluation, which aims to evaluate the acquisition of cultural knowledge (or of *a culture of individuals*), while on the other hand *procedural* evaluation focuses on the acquisition of discoursal and behavioural knowledge (or the acquisition of *an individual culture*). A number of evaluative procedures were tried out and concrete examples from the class studied are presented. The article concludes by examining the validity of the evaluative or auto-evaluative practices applied during the stay in the host country and discussing possible procedures for the analysis of what is effectively acquired culturally by the learner in terms of interpretative and behavioural competencies.

**01–273 Guerrero, Michael D.** (U. of Texas at Austin, USA; *Email*: mdguerrero@mail.utexas.edu). The unified validity of the Four Skills Exam: applying Messick's framework. *Language Testing* (London, UK), **17**, 4 (2000), 397–421.

In the USA, 17 states use Spanish-language proficiency tests to ensure that bilingual education teachers are able to deliver academic instruction in Spanish to school-age students. However, little is known about the tests' validity. In the study reported here, the unified validity of the Four Skills Exam (FSE), used in New Mexico for nearly 18 years, was evaluated using Messick's (1989) framework. Through this analysis, the FSE was found to lack an adequate degree of validity due to questionable psychometric properties and factors external to the test. It is suggested that Messick's framework offers bilingual educators a comprehensive blueprint for assessing the validity of these tests. The review of the FSE also elucidates pitfalls in bilingual education teacher language proficiency testing which other states should avoid.

**01–274 Han, Youngju** (Youngsan U.). Grammaticality judgement tests: how reliable and valid are they? *Applied Language Learning* (Presidio of Monterey, CA, USA), **11**, 1 (2000), 177–204.

A number of researchers now recognise that grammaticality judgement data do not always reflect linguistic knowledge and that they may lack reliability. The study reported here addresses the issue of the reliability of grammaticality judgement tests and explores what it is that they measure (i.e., their construct validity). Various methods of examining their reliability demonstrate that grammaticality judgement tests used in this study had relatively low reliability. The analyses of response patterns suggest some doubts about the extent to which grammaticality judgement data represent learners' grammatical knowledge. The weak relationship between timed and delayed judgements suggests that learners may use different types of knowledge under different task conditions. Qualitative analysis of the interview data indicates considerable confusion and indeterminacy in the learners' judgements. The results of the study suggest that researchers should be aware that there is a problem of reliability and validity in grammaticality judgement tests as an instrument for investigating learners' knowledge of grammatical rules.

**01–275 Kohonen, Viljo** (U. of Tampere, Finland). Student reflection in portfolio assessment: making language learning more visible. *Babylonia* (Comano, Switzerland), **1** (2000), 15–18.

The author maintains that skills-oriented, quantitative testing of linguistic proficiency tells only half the story, and that a number of important properties which are educationally valuable are 'invisible' in traditional assessment. Such factors as learners' willingness to take risks in order to cope with communicative tasks/situations and the skills and strategies they develop in order to foster independent, effective language learning apparently impinge directly on their observable language performance. The author groups language learning goals which aim to develop these skills under three main-categories: task, personality and process/context awareness, the latter (for example) involving how far learners are able to assess their language ability. It is claimed that a portfolio method of assessment is capable both of highlighting these learner characteristics, and enhancing them through explicit pedagogy. Using a European Language Portfolio model, the author describes a scheme in Finland where secondary school students kept a personal learning diary, identifying their strengths, weaknesses, and expectations of themselves and their teachers. Each piece of work submitted for the portfolio was also accompanied by formal student reflection in the diary on the learning process itself and on the learner's successes/failures. The article concludes that this process involves a

shift in the teacher's role to that of observer, enabler and organiser of learning opportunities.

**01–276 Langner, Michael** (Universität Freiburg, Germany). Online-Tests, ausprobiert! Was leisten Fremdsprachen-Tests im Internet? [Online tests tested! What do foreign language tests on the Internet achieve?] *Babylonia* (Comano, Switzerland), **1** (2000), 55–59

The author of this article briefly describes 16 language tests which he managed to locate (and subsequently try out) on the Internet. Most of them focus on English, although a small number are concerned with German, French and Spanish. The author draws out two central points. The first is that his results in the English tests diverged very widely. He also concludes that a comparison of the different tests suggested that none of them fully exploited the available multimedia resources.

**01–277 Lenz, Peter** (Universität Freiburg, Germany). Erfahrungen mit dem Europaischen Sprachenportfolio in der Schweiz. [The European Language Portfolio: the Swiss experience.] *Babylonia* (Comano, Switzerland), **1** (2000), 23–28.

The European Language Portfolio (ELP) has been tried out in various countries across Europe between 1998 and 2000. Switzerland has been one of the countries taking part, with more than 100 classes participating across all education sectors except primary. At the European level, the initiative has been co-ordinated and evaluated by the Council of Europe; at the local level, in Switzerland, by the CDIP (*Conférence des directeurs de l'instruction publique*). The provisional results of the Swiss evaluation indicate that the ELP is in principle considered by teachers to be an efficient and innovative tool which largely fulfils its intended brief. But its practical implementation in classrooms has sometimes proved difficult; and its format and ease of use have been open to criticism. Many, though not all, of the criticisms are in line with its experimental nature: the ELP has yet to make its mark, both in the education sector and the world of work. It is concluded that the European Year of Languages in 2001 will provide a good opportunity to promote an improved version of the ELP.

**01–278 Meara, Paul, Rodgers, Catherine and Jacobs, Gabriel** (U. of Wales Swansea, UK; *Email*: p.m.meara@swansea.ac.uk). Vocabulary and neural networks in the computational assessment of texts written by second-language learners. *System* (Oxford, UK), **28**, 3 (2000), 345–54.

This paper explores the potential of a neural network in language assessment. Many examination systems rely on subjective judgements made by examiners as a way of grading the writing of non-native speakers. Some research (e.g., Engber, 1995, The relationship of lexical proficiency to the quality of ESL compositions, *Journal of Second Language Writing*, 4, 2, 139–55) has shown that these subjective judgements are influenced to a very large extent by the lexical choices made by candidates. The present authors took Engber's basic model, but automated the evaluation of lexical content. A group of non-native speakers of French were asked to produce a short text in response to a picture stimulus. The texts were graded by French native speaker teachers. A number of words which occurred in about half the texts were identified, and each text was coded for the occurrence and non-occurrence of each word. A neural network was then trained to grade the texts on the basis of these codings. The results suggest that it might be possible to teach a neural network to mimic the judgements made by human markers.

**01–279 North, Brian** (Eurocentres, Zurich). Adaptive testing. *Babylonia* (Comano, Switzerland), **1** (2000), 50–54.

This article presents adaptive testing as a quick, reliable and accurate way of assessing language proficiency. Traditional tests, in contrast, are seen as context-dependent, reliant on 'expert' judgement, and difficult to compare. The pros and cons of four test design options are discussed. The first, 'start at the bottom and climb', involves a large number of discrete-point test items at each level of the scale, usually in ascending order of difficulty – such tests are time-consuming, and mid-ability candidates waste time completing those which are too easy and tackling others which are too difficult. The 'fast track/fine-tuning' procedure (e.g., Test of Proficiency in English and the Eurocentre's Vocabulary Size test) involves 'branching', where success on earlier parts of the test channels candidates to targeted tasks at the appropriate difficulty level. This flexibility, however, can be compromised if candidates are nervous and perform badly at the outset. Adaptive testing uses items which have a known 'difficulty value'; the computer analyses candidates' responses to each item and assigns an 'ability value', on the basis of which their proficiency level can be mapped. The article concludes by describing five CATs (Computer-Adaptive Tests), including Eurocentre's Itembanker, and CBIELTS (a planned computer version of the IELTS test). It is claimed that CATs are most appropriate with items attached to relatively short, screen-sized passages rather than longer academic texts and the comprehension/inferencing items typically associated with them.

**01–280 Oller, Jr., John W.** (U. of Louisiana at Lafayette, USA; *Email*: joller@louisiana.edu), **Kim, Kunok and Choe, Yongjae**. Applying general sign theory to testing language (verbal) and nonverbal abilities. *Language Testing* (London, UK), **17**, 4 (2000), 377–96.

This article presents the basis for a general theory of signs showing that the relation between acquired language proficiencies and so-called nonverbal abilities must be closer than has been commonly supposed. A

general theory of signs shows a deep logical dependency of nonverbal/performance tests on conventional linguistic signs. Two hypotheses follow: (1) to the extent that nonverbal abilities can be measured, they must be positively correlated with primary language abilities; (2) (a) in the early stages of acquiring a second or foreign language, proficiencies in the primary (stronger or native) language should correlate more strongly with nonverbal abilities than proficiencies in the nonprimary (weaker, second or foreign) language, and (b) as persons approach greater parity between their primary and any nonprimary language, correlations between nonverbal scores and proficiencies in the two languages should both be significantly positive, and should approach equality. By contrast, the Cattell-Horn theory predicts about the same level of correlation throughout the course of development, and Gardner's theory of multiple intelligences predicts that distinct intelligences should be uncorrelated. The present authors suggest that all three theories can be tested by examining the simple correlations of language proficiency measures with nonverbal IQ scores in intermediate and advanced nonprimary language learners.

**01–281 Oscarson, Mats** (Göteborg U., Sweden). Selbstbeurteilung im Fremdsprachenunterricht – eine Utopie? [Self-assessment in foreign language learning – a Utopian vision?] *Babylonia* (Comano, Switzerland), **1** (2000), 19–22.

In current practices in language education students are often more closely involved in the evaluation of their own learning than previously. This article sketches out the background to this development and makes a case for learner self-assessment. It is pointed out, for instance, that a great deal of language learning takes place outside the classroom (notably in the most widely taught language, i.e., English) and that the opportunities for realistic self-testing have thereby increased considerably. Indeed, it is argued that certain aspects of communicative ability are very difficult to assess outside real-life language use settings (e.g., in the classroom). Self-monitoring ability also makes independent learning after formal schooling easier and more effective. A number of investigations of the validity of self-ratings are reviewed. Results obtained have been encouraging in most cases. Learners tend to have a fairly good grasp of their ability, although some researchers have also reported negative outcomes. It is concluded that self-assessment is *not* a Utopian vision: it has been shown to work in many contexts and with a variety of learners; and there is also some evidence that the necessary skills can be learnt. Accustoming learners, through training, to the idea of self-managed assessment is in any case recommended.

**01–282 O'Sullivan, Barry** (U. of Reading, UK; *Email*: b.e.osullivan@reading.ac.uk). Exploring gender and oral proficiency interview performance. *System* (Oxford, UK), **28**, 3 (2000), 373–86.

There is growing interest in those factors which affect the test performance of the language learner, some of it motivated by a desire to detect and eliminate test features which are seen as distorting the tester's attemps to achieve accurate assessment of learners' language proficiency. A number of researchers, however, distinguish between test features which are irrelevant to the ability which is being measured, and those which are relevant. It is important to discover which test features constitute significant sources of true variance in leaners' performance. One feature which has been shown to affect learners' performance on tests of spoken interaction is the gender of the person with whom they interact. This article reports a study in which 12 Japanese learners were interviewed, once by a man and once by a woman. Videotapes of these interactions were scored by trained examiners. Comparison of scores awarded indicated that in all but one case the learners performed better when interviewed by a woman, regardless of the sex of the learner. Sixteen interactions, involving eight learners, were then transcribed. Analysis of interviewer-language indicated systematic gender differences, while analysis of the responses of the learners suggest a tendency to produce more grammatically accurate language with their female interviewers.

**01–283 Stotz, Daniel** (Hochschule Winterthur, Switzerland). Evaluation im frühen teilimmersiven Fremdsprachenunterricht. [Evaluation in early part-immersion language learning.] *Babylonia* (Comano, Switzerland), **1** (2000), 29–32.

This article reflects on the assessment of language competence of young learners who are learning English from the first grade of primary school in an experimental approach termed School Project 21 in the Canton of Zurich. It argues that in a partial immersion programme, here referred to as *Embedding*, assessment procedures must be derived from a close observation of classroom teaching and learning in order to cope with the high degree of context variability and heterogeneous perceptions of objectives. An approach to assessment is postulated which reflects the embedded teaching and learning and is capable of delivering dense description. It is suggested that an adaptation of the Council of Europe reference framework to primary school contexts may go some way towards telling teachers, parents and taxpayers how well the children master the language to which they are exposed from an early age.

# Teacher education

**01–284 Boyle, Joseph** (The Chinese U. of Hong Kong, China). Education for teachers of English in China. *Journal of Education for Teaching* (Abingdon, UK), **26**, 2 (2000), 147–55.

This article outlines some of the ways in which foreign teachers who wish to teach English in modern China