back messages to guide learners to correct their errors by themselves. Another system (called the M system) enables learners to input their answer from a multiple selection, and displays only the correct answer as feedback, regardless of the learner's response. Twenty-two Japanese language learners were involved in a comparison experiment to examine the usefulness of the T system. The results show that the T system is preferred to the M system in the input method and the feedback method. The results indicate that the 'freely input' method and the 'feedback corresponding to the learner's typed sentence' is better than the 'multiple selection' method and 'feedback that only displays the correct answer', respectively.

Language testing

00–183 Barnes, Ann, Hunt, Marilyn and Powell, Bob (U. of Warwick, UK). Dictionary use in the teaching and examining of MFLs at GCSE. *Language Learning Journal* (Rugby, UK), **19** (1999), 19–27.

This article considers the recent introduction of bilingual dictionaries into examinations in modern foreign languages (MFLs) in England and Wales for the General Certificate of Secondary Education (GCSE) and the implications for both teachers and learners. It discusses the context and describes in detail teachers' perceptions of this development through the analysis of data obtained from two small-scale questionnaire surveys. Responses from MFLs teachers in 100 secondary schools suggest that, on the whole, they are positive about the use of dictionaries, particularly when one considers the major shift of policy in recent years: they appear to welcome the introduction of dictionaries and to have considered the surrounding issues very carefully. The article concludes, however, that there is a need for more research into the use, effects and integration of dictionaries in the MFLs curriculum at this level, particularly the impact on pupils' performance in examinations.

00–184 Brown, James D. and Hudson, Thom (U. of Hawai'i, USA). The alternatives in language assessment. *TESOL Quarterly* (Alexandria, VA, USA), **32**, 4 (1998), 653–75.

This article posits that language testing differs from testing in other content areas because language teachers have more choices to make. The authors set out to help language teachers to decide what types of language tests to use in their particular institutions and classrooms for their specific purposes. The various kinds of language assessments are classified into three broad categories: (a) selected-response assessments (including true-false, matching, and multiple-choice); (b) constructed-response assessments (including fill-in, short-answer, and performance); and (c) personal-response assessments (including conference, portfolio, and self- or peer-assessments). A clear definition is provided for each assessment type, and the advantages and disadvantages of each are explored. The article concludes with a discussion of how teachers can make rational choices among the various assessment options by thinking about: (a) the consequences of the washback effect of assessment procedures on language teaching and learning; (b) the significance of feedback based on the assessment results; and (c) the importance of using multiple sources of information in making decisions based on assessment information.

00–185 Buckby, Mike. The use of the target language at GCSE. *Language Learning Journal* (Rugby, UK), **19** (1999), 4–11.

The project reported in this article aimed to evaluate whether GCSE (General Certificate of Secondary Education) examinations in England and Wales with a greatly increased role for the target language maintain the standards of previous examinations. To this end, an experimental examination based on items from past GCSE examinations was devised and administered to 151 Year 11 students (aged 15–16). The article reports the examination and the results in some detail. Results suggested that (a) the demands of target language examinations *are* in line with past demands, and (b) a hierarchy of question types is needed to enable candidates to demonstrate their competencies.

00–186 Chalhoub-Deville, Micheline and Deville, Craig (U. of Iowa, USA). Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics* (Cambridge, UK), **19** (1999), 273–99.

This article lists the advantages and drawbacks of Computer Adaptive (CAT) and Computer-Based Testing (CBT), and also outlines various projects wherein this technology is being developed. Despite advances in computer technology and measurement theory, computer-based testing still basically focuses on selected response, discrete point tasks rather than performance-based items. Immediate feedback can be provided with regard to a learner's total score, and the use of Item Selection Algorithms allows each testee to take a different path through the test, as the items selected for completion depend on how successful he/she has been in dealing with previous exercises. In effect, CBT comprises 'branching programs' which tailor the item repertoire/structure to the performance profile of the candidate, as it develops. A potential problem, however, is the omission of essays and interviews, for example, which may restrict CAT to testing linguistic knowledge rather than communicative skills. The authors outline important CAT design issues (such as ensuring the provision of a large enough item pool, the need to document comparability of scores between computer adaptive tests and their P&P 'pencil and paper' counterparts, and determining the appropriate entry/exit points in the test), and also refer to innovations such as COMPASS (a written ability test) and the comput-

erised version of TOEFL (Test of English as a Foreign Language). Consideration is given to CAT projects in different languages (e.g English, French, Dutch, Hausa) being developed in the United States and Europe (the authors discuss the UCLES 'CommuniCAT'), and it is concluded that although CBT has developed significantly over the past decade it is not a testing panacea and must be viewed with its current limitations (particularly expense and technological complexity) in mind.

00–187 Chapelle, Carol A. (Iowa State U., USA). Validity in language assessment. *Annual Review of Applied Linguistics* (Cambridge, UK), **19** (1999), 254–72.

This review paper first briefly outlines the history of validation in language testing, including an annotated bibliography. It then compares traditional views–where validity is perceived as an internal characteristic of language tests–with more recent perspectives which emphasise validity as pertaining to score interpretation, as well as the uses to which these are put and the attendant consequences (e.g., in making university admission decisions on the basis of test results). Contemporary testing philosophy also apparently perceives construct validity as a unitary concept, with correlational and content factors (seen previously as separate types of validity in their own right) in subsidiary roles. The paper considers key areas currently debated, e.g., the role of hypotheses in evaluating testing outcomes and the identification of relevant evidence for testing validation hypotheses. Six approaches to the provision of validity evidence (notably, analysis of test content, items/tasks, 'dimensionality' and testing consequences) are also summarised. Perceived challenges relating to the current, broadened view of test validation include the development of a coherent construct theory upon which hypotheses can be formulated and against which test data can be evaluated. Validation is seen as expensive and time-consuming, and in practical terms, context-specific rather than definitive, proof-based or generalisable. The greatest challenge is seen as adapting the current literature on validity into actual practice in second language classes, programmes and research.

00–188 Chiang, Steve Y. (Da Yeh U., Changhua, Taiwan). Assessing grammatical and textual features in L2 writing samples: the case of French as a foreign language. *The Modern Language Journal* (Malden, MA, USA), **83**, 2 (1999), 219–32.

This article investigates the relative importance of various grammatical and discourse features in the evaluation of second language (L2) writing samples produced by college students enrolled in beginning and intermediate French courses. Three native-speaking instructors of French rated 172 essays using a scale constructed by the researcher and based on theory and research from discourse analysis. The scale contained four areas of evaluation–morphology, syntax, cohesion and coherence–encompassing a total of 35 language/textual features, in addition to a holistic judgement of overall quality. Among the findings are that (a) raters relied heavily on discourse features, especially those for cohesion, in judging the overall quality of an essay; and (b) the rating scale exhibits content validity and reliability, although refinement is still needed to achieve a desired construct validity. It is suggested that future research should focus on discovering other elements involved in the rating practice through analytical delineations and validation procedures and on adapting the proposed rating instrument for large-scale assessment contexts.

00–189 Fulcher, Glenn (U. of Surrey, UK). Assessment in English for Academic Purposes: putting content validity in its place. *Applied Linguistics* (Oxford, UK), **20**, 2 (1999), 221–36.

Testing and assessment in English for Academic Purposes (EAP) contexts has traditionally been carried out on the basis of a needs analysis of learners or a content analysis of courses. The present author deems this unsurprising, given the dominance of needs analysis models in EAP, and a focus in test design that values adequacy of sampling as a major criterion in assessing the validity of an assessment procedure. This article reassesses this approach to the development and validation of EAP tests on the basis of the theoretical model of Messick (1989) and recent research into content specificity, arguing that using content validity as a major criterion in test design and evaluation has been mistaken.

00–190 Kunnan, Antony J. (California State U., USA). Recent developments in language testing. *Annual Review of Applied Linguistics* (Cambridge, UK), **19** (1999), 235–53.

This review article discusses six new developments in language testing: the role of ethics, the expanded view of validation and the role of fairness, applications of Structural Equation Modelling, the Computer-Based TOEFL (Test of English as a Foreign Language), the TOEFL 2000 project, and recent additions to the UCLES EFL examinations suite. Testing issues outside the UK and the USA are also considered, and a selective overview of new resources is presented, including *Language Testing and Assessment* (Clapham & Corson (eds.), 1997), the Association of Language Testers in Europe's multilingual glossary of terminology, and the *Dictionary of Language Testing* (Davies, Brown, Elder, Hill, Lumley & McNamara, in press). Considering papers and publications from various symposia and conferences, the article also highlights two small but significant research projects: the evaluation of *PhonePass* (an English Second Language speaking test administered by telephone) using Bachman and Palmer's six test criteria (1996), and an investigation into the appropriacy/practicality of NNS (Non-Native Speaker) English language admissions criteria at the University of Lancaster. The author concludes that testing now clearly comprises more than quantitative methodology/data analysis, and that the exploration of issues with a social dimension (e.g., fairness and validation) indicates an important broadening of the field. An annotated bibliography is included.

00–191 Pintori, Adriana (U. Autònoma de Barcelona, Spain). Indici di frequenza di errori nella prova di comprensione dell'italiano come L2. [Error frequency indices in the comprehension test of Italian as a second language.] *Rassegna Italiana di Linguistica Applicata* (Rome, Italy), **1** (1999), 37–47.

First language interference is a complex phenomenon, whose effects are especially salient in the case of closely related languages. The author of this brief report illustrates the contrastive implications of a written comprehension test administered to Spanish undergraduates learning intermediate-level Italian at a school for translators and interpreters. Thirty-two students were given a full-length newspaper article and asked to answer 20 multiple-choice questions, consisting mostly of word definitions/synonyms and a few grammar transformation points: as expected, the items targeting likely sources of interlinguistic error proved particularly difficult. It is concluded that, in mixed language classrooms, separate versions of the same test should therefore be used to ensure that no particular language group is unduly favoured or penalised.

00–192 Rossiter, Marian and Pawlikowska-Smith, Grazyna (U. of Alberta, Canada). The use of CLBA scores in LINC program placement practices in Western Canada. *TESL Canada Journal / La Revue TESL du Canada* (Burnaby, B.C.), **16**, 2 (1999), 39–52.

The *Canadian Language Benchmarks Assessment* (CLBA) tool aims at helping to place language learners across Canada in instructional programmes appropriate to their level of competence in English. Introduced in 1997, the CLBA, a low-stakes task-based test, measures the lower eight of the 12 levels described by the *Canadian Language Benchmarks*. The CLBA is administered to adult immigrants to Canada who are eligible for Language Instruction for Newcomers to Canada (LINC), a programme funded by the federal government. This article reports the results of a survey of 19 ESL (English Second Language)/LINC programmes in Western Canada regarding the use of CLBA scores for in-class placement purposes. It is suggested that, in its present form, the CLBA appears to have only limited usefulness as an in-class placement instrument: it is criterion-referenced, and reports broad-band proficiency information, but does not discern small differences between individuals' performances. The authors present and discuss such factors contributing to the need, where applicable, for additional placement testing by some programmes, and how these factors are in part related to different functions of tests, to varying approaches to scoring (holistic and analytical), and to the format by which the assessment results are communicated. Issues of concern to LINC administrators and recommendations for addressing these issues are also presented.

00–193 Hartley, Linda and Spöring, Marion (U. of Dundee, Scotland). Teaching communicatively; assessing communicatively? *Language Learning Journal* (Rugby, UK), **19** (1999), 73–79.

This article reiterates the case for the teaching of communicative competence, which it defines (after Canale and Swain, 1980) in terms of grammatical, sociolinguistic, discourse and strategic competencies. It argues that the only appropriate assessment for communicative teaching is the assessment of communicative performance in a contextualised situation, under realistic linguistic, situational, cultural and affective constraints. It goes on to illustrate a range of integrative, task-based assessments which have been successfully used within the authors' own institution, and shows how these may be categorised according to stage or level of study, text type (of test prompt), task type, linguistic skills used, language used (first or target) and type of answer (in terms of tenor and mode). The authors consider that such communicative assessments are no less reliable than traditional assessments and that their validity may well be higher.

00–194 Taylor, Carol (Educational Testing Service, Princeton, USA), **Kirsch, Irwin, Eignor, Daniel and Jamieson, Joan**. Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning* (Malden, MA, USA), **49**, 2 (1999), 219–74.

The planned introduction of a computer-based Test of English as a Foreign Language (TOEFL) test raises concerns that language proficiency will be confounded with computer proficiency, introducing construct-irrelevant variance to the measurement of examinees' English language abilities. In the study reported here, a questionnaire focusing on examinees' computer familiarity was administered to 90,000 TOEFL test takers. A group of 1,200 'low-computer-familiar' and 'high-computer-familiar' examinees from 12 international sites worked through a computer tutorial and a set of 60 computer-based TOEFL test items. No meaningful relationship was found between level of computer familiarity and level of performance on the computerised language tasks after controlling for English language ability. It was concluded that no evidence exists of an adverse relationship between computer familiarity and computer-based TOEFL test performance due to lack of prior computer experience.

00–195 Walsh, Polly (Universita degli Studi di Firenze, Italy). Can a language interview be used to measure interactional skill? *CALS Working Papers in TEFL* (U. of Reading, UK), **2** (1999), 1–28.

Whilst recognising that a speaking test can never resemble a spontaneous conversation, this paper argues that it may nonetheless share features with other real-life forms of asymmetrical, collaborative discourse, including other types of interview and naturalistic native speaker-nonnative speaker (NS-NNS) exchanges. Using analytical frameworks derived from studies of conversation man-

agement, it focuses on the relative linguistic proficiency, and hence interactional status, of the participants, arguing that the closer they come to being linguistic equals, the more likely they are to collaboratively construct the discourse. The paper goes on to report a small-scale study of language interviews with adult students in an Italian university, focusing on those features which appeared to vary according to the relative proficiency of the participants, including: amount of talk; topic initiations and prompts; interruptions, overlaps and back-channelling; and linguistic accommodation (on the part of the more proficient speaker). It suggests that a fuller understanding of the patterns of interaction in naturalistic NS-NNS discourse could lead to the design of speaking tests and rating scales which would more genuinely reflect communicative competence.

00–196 Welling-Slootmaekers, Margriet (Cito–Nat. Inst. for Educational Measurement, Arnhem, The Netherlands). Talenexamens vbo, mavo en havo vanaf 2000. De belangrijkste veranderingen. [Language examinations in Dutch secondary school from 2000 onward. The main changes.] *Levende Talen* (Amsterdam, The Netherlands), **542** (1999), 488–90.

For several decades, all national language examinations in Dutch secondary education were reading comprehension tests consisting entirely of multiple-choice questions. From 2000 onward, the examinations will also include a number of open-ended questions, which the pupils are supposed to answer in their mother tongue, Dutch. The author argues that this change will not create any problems, as the pupils are used to open-ended questions in all other national examinations; language examinations have been an exception in this respect. In terms of rating the examinations, the change implies that teachers will have to rate their pupils' examinations and that a second judge will be required, as open-ended questions cannot be rated automatically. The author presents examples of the new question types, which include: short-answer questions, quotation questions (e.g. 'Which words in the 4th paragraph indicate that X is mad at Y?'), completion questions, regrouping questions and long-answer questions. Cito has published sample examinations for English, French, German and Spanish.

Teacher education

00–197 Asghar, Saima Ali (U. of Warwick, UK). Staff appraisal in education: perceptions and practices across cultures. *English Language Teacher Education and Development (ELTED)* (U. of Warwick / U. of Birmingham, UK), **4**, 1 (1998), 47–65.

This article reports a study conducted at the author's institution which looks at the varied styles of staff appraisal experienced by the participants in the study–a group of twelve teacher trainees and teaching staff–along with their perceptions of good staff appraisal procedures and practices. It is what Holliday (1997) calls 'an ethnographic study' from 'varied and locationally spread' teaching environments around the world. It is motivated by the views expressed by the author's colleagues on the MA programme which highlighted the fact that there are radically different perspectives on teacher development, and the realisation that there is a need to bring some coherence to these potentially conflicting viewpoints. The research gives a glimpse into the diverse appraisal processes experienced by the multi-cultural group of participants by eliciting their views of their appraisal experiences, and the combinations of ways in which appraisal is conducted across the world. This has led to the conclusion that, while most innovations in teaching and staff development stem from 'BANA' (Britain, North America, Australasia) cultures as identified by Holliday (1993), they need to be adapted to suit the culture of the institution that they are exported to in the rest of the world.

99–198 Bucher-Poteaux , Nicole (U. Louis Pasteur, Strasbourg, France). Les mémoires professionels à l'IUFM. [Professional reports in University Teacher Training Institutes]. *Les Langues Modernes* (Paris, France), **1**, (1999), 6–7.

This paper introduces an issue dedicated to discussion of the professional reports required of trainees during their second year at the IUFM. The reports, which require the linking of theory and practice, represent an innovation in the assessment of aspiring teachers and complement other, more theoretical assessment processes. Five of the seven papers have been written by IUFM teaching staff. **Clerc** examines the role of writing in a teacher's life, noting the contrast between the mainly oral life of the classroom and staffroom and the reflective demands of the written mode. She argues strongly for the importance of professional writing in the development of practical competence. **Tournadre** confronts the difficulties of the professional report (often seen as a relic of the academic tradition; an unwelcome pressure added to that of teaching practice, leading frequently to unsatisfactory pieces of work), but concludes that there is much of value in the work required to complete the report and offers constructive suggestions for the improved presentation and justification of the task to the trainees. **Harambour** analyses the topics of reports written by English trainees over a four-year period. These reveal a tendency to generalisation rather than application to a particular situation. As a result, ready-made solutions, sometimes inappropriate to the context, are chosen. A complementary survey of trainees' perceptions showed that negative attitudes to the task were often associated with a lack of understanding of its objectives. More positively, **Moll** explores the role of the report supervisor, valuing the supportive relationship which can be developed with the trainee when both regard the task as one of exploration. In two papers