has been suggested that solving an essay question requires the activation of processes similar to those involved in solving a physics problem, a cognitive-metacognitive framework developed for mathematical problem solving was applied to essay writing. To test this hypothesis, a small number of L2 students of various abilities were asked to write an essay while thinking aloud, and their protocols were analysed in the light of the mathematical framework. Results indicate that the framework is a reliable tool provided that certain adjustments are made. These adjustments are discussed, and a cognitive-metacognitive model more geared towards essay writing is suggested.

99–621 Lock, Graham and Lockhart, Charles (City U. of Hong Kong). Genres in an academic writing class. *Hong Kong Journal of Applied Linguistics* (Hong Kong), **3**, 2 (1998), 47–64.

This paper identifies and describes the genres that a group of tertiary level English Second Language students produced during a process writing class in which they were free to decide their own topics, purposes and audiences. Participants were 27 students randomly selected from 54 Cantonese-speaking first-year university students. Six expository genres are identified: description, advice, analysis, report, discussion, and argument. Characteristics of these genres, the relationships among them, and their schematic structures are described. It is concluded that students in the study simply reproduced the genres familiar from secondary school and its examination-oriented syllabus. It is argued that students need to extend their repertoire of genres beyond those of their previous educational experience by writing on more specialised topics for more specific audiences and purposes, and by producing longer and more complex texts.

99–622 New, Elizabeth (U. of North Texas, USA). Computer-aided writing in French as a foreign language: a qualitative and quantitative look at the process of revision. *The Modern Language Journal* (Malden, MA, USA), **83**, 1 (1999), 80–97.

Little documentation currently exists on the writing strategies and habits of foreign language (FL) writers. This study was designed to observe, as unobtrusively as possible, the revision strategies of five students of FL French enrolled in a one-semester intensive intermediate college French course. The participants completed a two-part writing task with the aid of the software program *Système-D* (Noblitt, Solá & Pet, 1987, 1992). Of considerable interest is the program's keystroke tracking device, which records the lexical, grammatical and thematic information accessed by students while writing. Analysis of the compositions, computer records, videotapes of writing sessions, and student responses to post-writing questionnaires provide a detailed picture of how and when the students revised in real time – with minimal impact on the writing process itself. Results showed that both the self-reported good writers and poor writers engaged in the process of revising and that, as expected, surface-level changes far outnumbered the changes to content. These findings suggest that linguistic concerns and lack of explicit instruction on revision and computer strategies impede the reviewing and reworking of texts.

99–623 Parks, Susan (Université Laval, Quebec, Canada) **and Maguire, Mary H.**. Coping with on-the-job writing in ESL: a constructivist-semiotic perspective. *Language Learning* (Malden, MA, USA), **49**, 1 (1999), 143–75.

Despite a long-standing interest within applied linguistics in the analysis of written genres, few studies have attempted to show how such genres are appropriated by new members in second language (L2) academic or workplace settings. Based on a 22-month qualitative study, this article reports on how francophone nurses, who were newly hired in an English-medium hospital in Montreal, Canada, developed skill in writing nursing notes – which differed from the way they were done in French – in English. Central to the analysis is the construct of mediation, explored in terms of how collaborative processes, both overt and covert, shape text production as well as other less visible, taken-for-granted aspects of the social context. The article concludes with reflections on the implications of such inquiries for L2 writing theory, research, and practice.

99–624 Ramanathan, Vai and Atkinson, Dwight (U. of Alabama, USA). Ethnographic approaches and methods in L2 writing research: a critical guide and review. *Applied Linguistics* (Oxford, UK), **20**, 1 (1999), 44–70.

This paper discusses central concepts and issues regarding ethnographic research in education, particularly as they pertain to studies of second language (L2) writing. After a consideration of Watson-Gegeo's (1988) six principles of ethnographic research, the present authors propose their own 'prototype' definition. Following a discussion of some key concepts in that definition, they then review three recent studies of L2 writing which are ethnographic in nature. They next discuss the vexed issue of 'generalisability', and consider two further studies of L2 writing in that regard. They end by introducing a series of issues which are critical to recent ethnographic concerns in anthropology and sociology, but which have had little influence so far on ethnographically oriented L2 writing research.

Language testing

99–625 Beglar, David (Temple U., Osaka, Japan) **and Hunt, Alan**. Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing* (London, UK), **16**, 2 (1999), 131–62.

Few researchers have undertaken detailed investigations of the reliability and validity of tests designed to measure vocabulary size. The purpose of this study, there-

fore, was to carefully analyse and validate the revised versions of the 2000 Word Level and the University Word Level of Nation's Vocabulary Levels Test (Nation 1990), which can be administered for the purposes of course planning and placement in language programmes. In this study, 496 Japanese high school and university students completed four forms of the 2000 Word Level Test, and 464 participants completed four forms of the University Word Level Test. Two new forms of each test were created and then examined using classical and Rasch item analyses. The new forms were found to be acceptably reliable, and the Rasch analysis revaled only three misfitting items. In addition, these forms had statistically significant correlations with the TOEFL (Test of English as a Foreign Language), particularly for the reading and grammar sub-sections. A variety of evidence is then presented to support the notion that the tests are unidimensional. Finally, the authors suggest further improvements and new directions for future research.

99–626 Brown, James Dean (U. of Hawai'i, Manoa). The relative importance of persons, items, subtests and languages to TOEFL test variance. *Language Testing* (London, UK), **16**, 2 (1999), 217–38.

The purpose of this project was to explore the relative contributions to TOEFL (Test of English as a Foreign Language) score dependability (which is analogous to classical theory reliability) of various numbers of persons, items, subtests, languages and their various interactions. Three research questions were formulated: (1) what the characteristics of the distributions are, and how high the classical theory reliability estimates are for the whole test and its subtests; (2) for each of the 15 languages, what the relative contributions to test variance are of persons, items, subtests and their interactions; and (3) across all 15 languages, what the relative contributions to test variance are of persons, items, subtests and languages, as well as their various interactions. The study sampled 15,000 test takers, 1000 each from 15 different language backgrounds, from the total of 24,500 participants in the TOEFL generic data set which itself was a sample from the May 1991 worldwide administration of the TOEFL. The test, administered under normal operational conditions, included all three subtests: Listening Comprehension, Structure and Written Expression, and Vocabulary and Reading Comprehension. The analyses included descriptive statistics, classical theory reliability estimates, and a series of generalisability studies conducted to isolate the variance components due to persons, items, subtests and languages, and their effects on the dependability of the test. Unlike previous research, the results here indicate that, when considered in concert with other important sources of variance (persons, items and subtests), language differences alone account for only a very small proportion of TOEFL test variance.

99–627 Bhgel, Karin and Leijn Melse (Cito Arnhem, The Netherlands). Nieuwe examens havo/vwo, nieuwe vraagvormen. Onderzoek naar de objectiviteit van de beoordeling van open vragen moderne vreemde talen. [New exams in secondary education, new question types. An investigation into the reliability of the evaluation of open-ended questions in foreign-language exams.] *Levende Talen* (Amsterdam, The Netherlands), **537** (1999), 173–81.

This paper deals with foreign-language exams in Dutch secondary education. Since the early 70s, these exams consist of reading comprehension tests that employ only multiple-choice questions. The new exam regulations allow open-ended questions for one-third of the exam. Although open-ended questions have many advantages – not least in terms of face validity – they are also problematic in a number of respects. An investigation among a random sample of teachers and pupils was carried out. Despite the fact that only short-answer questions were used and that judges were given an evaluation model, the results showed that inter-rater reliability was below standard (.65 on a scale of 0 to 1). The authors recommend a re-evaluation of existing exam practices in at least two respects: exams should be judged anonymously and by two independent judges.

99–628 Coniam, D. (The Chinese U. of Hong Kong). Voice recognition software accuracy with second language speakers of English. *System* (Oxford, UK), **27**, 1 (1999), 49–64.

This paper explores the potential of the use of voice recognition technology with second language (L2) speakers of English. Developing his earlier study with a small group of native speakers, the author analyses the output produced by very competent L2 participants reading a text into the voice recognition software Dragon Systems *Dragon NaturallySpeaking*. As the program is speaker-dependent and has to be trained to recognise each person's voice, participants first spent about 45 minutes reading a training text of some 3800 words. They then read a second test text of some 1050 words. The output produced by the software was analysed in terms of words, sub-clausal units, clauses and t-units. In terms of accuracy, the L2 speakers' output was significantly lower than that achieved by the native speakers. The results were nonetheless consistent in line with the native speakers' scores, i.e., the highest accuracy scores were achieved at the lowest (and most discrete) level of analysis, the word level, and the lowest scores at the t-unit, or sentence level of analysis. The paper concludes that, although the technology is still at an early stage of development in terms of accuracy and single-speaker dependency, the consistent results achieved suggest that the development of an assessment tool, e.g., a reading-aloud test via voice recognition technology and determining a score through an analysis of the output, may be a testing procedure with potential.

99–629 Freedle, Roy and Kostin, Irene (Educational Testing Service, Princeton, NJ, USA).

Does the text matter in a multiple-choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* (London, UK), **16**, 1, 2–32.

This study addresses a specific construct validity issue regarding multiple-choice language-comprehension tests by focusing on TOEFL's (Test of English as a Foreign Language) minitalk passages, and asking whether there is evidence that examinees attend to the text passages in answering the test items. The authors analysed a large sample (n = 337) of minitalk items. The content and structure of the items and their associated text passages were represented by a set of predictor variables that included a wide variety of text and item characteristics identified from the experimental language-comprehension literature. Stepwise and hierarchical regression techniques showed that at least 33% of the item difficulty variance could be accounted for primarily by variables that reflected the content and structure of the whole passage and/or selected portions of the passage; item characteristics, however, accounted for very little of the variance. The pattern of these results was interpreted, with qualifications, as favouring the construct validity of TOEFL's minitalks. The methodology used also allowed a detailed comparison between TOEFL reading and listening (minitalk) items. Several criticisms concerning multiple-choice language-comprehension tests were addressed. Future work is also suggested.

99–630 Jafarpur, A. (Shiraz U., Iran). Can the C-test be improved with classical item analysis? *System* (Oxford, UK), **27**, 1 (1999), 79–89.

The application of the rule-of-'two' for constructing C-tests (i.e., half of the letters of every second word are deleted) produces two sorts of test items. Many items delineate acceptable facility and discrimination values, but a sizeable number of them are either extremely easy or extremely difficult to fill in. This paper describes a study undertaken in order to investigate whether this defect can be avoided. The study was designed to explore the effect of selecting C-test items on the basis of the statistical properties of individual items, and to investigate the extent to which the C-principle can be improved through classical item analysis. A C-test with five texts and 126 items was constructed and tried with 146 Iranian English majors. On the basis of an item analysis, a tailored C-test with 100 items was developed and tried with 60 other participants. The results of the study showed that no gains were made with the classical item analysis.

99–631 Jamieson, Joan (Northern Arizona U., USA), **Taylor, Carol, Kirsch, Irwin and Eignor, Dan**. Design and evaluation of a computer-based TOEFL tutorial. *System* (Oxford, UK), **26**, 4 (1998), 485–513.

In order to train examinees whose native language is not English to take a computerised Test of English as a Foreign Language (TOEFL), a special set of tutorials was designed and developed. These tutorials were trialled as part of a computer familiarity study in 1996; and this article describes their development. Also, the experiences of the 1169 participants in the study are characterised in terms of timing and performance data, as well as self-reported attitudes. These analyses took into account computer familiarity and English ability, which both proved to be important in explaining some differences in time to complete the tutorials and perception of the tutorials' usefulness. Most examinees were successful in completing the practice items in the tutorials and thought the tutorials helpful. Some changes were subsequently made before operational implementation of the computerised TOEFL test in order to reduce the time needed to complete the TOEFL tutorials.

99–632 Kormos, Judit (Eötvös U., Budapest, Hungary). Simulating conversations in oral-proficiency assessment: a conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing* (London, UK), **16**, 2 (1999), 163–88.

Several recent studies have investigated the nature of interaction in oral proficiency exams and have concluded that the interview format obscures differences in the conversational competence of the candidates. This paper examines the opportunities test takers have to display their knowledge of managing conversations in a second language in two types of task: non-scripted interviews and guided role-play activities. The data consist of 30 interviews and 30 role-play activities between near-native examiners and intermediate learners used in language exams in Hungary; they were analysed for the number of topics introduced and ratified by the examiner and the candidate respectively, as well as for the number of interruptions, openings and closings produced by them. The findings show that the conversational interaction is more symmetrical in the guided role-play activity, with the candidates introducing and ratifying approximately the same number of topics as the examiners. In addition, the examinees have the opportunity to interrupt and hold the floor more effectively and can demonstrate their knowledge of how to open and close a conversation. These findings suggest that guided role-play activities used in the study exhibit several characteristics of real-life conversations, and can therefore be used for assessing the candidates' conversational competence.

99–633 Laufer, Batia (U. of Haifa, Israel) **and Nation, Paul** (Victoria U. of Wellington, New Zealand). A vocabulary-size test of controlled productive ability. *Language Testing* (London, UK), **16**, 1, 33–51.

It is important in the design of the vocabulary component of a teaching programme that teachers are able to discover the state of their learners' vocabulary knowledge. It is also important that researchers can draw on a

variety of vocabulary measures to investigate the nature of vocabulary growth. The study reported here focuses on a controlled production measure of vocabulary consisting of items from five frequency levels, and using a completion item type such as *The garden was full of fra – flowers*. The controlled-production vocabulary-levels test was found to be reliable, valid – in that the levels distinguished between different proficiency groups – and practical. There was a satisfactory degree of equivalence between two equivalent forms of the test.

99–634 Papajohn, Dean (U. of Illinois, Urbana-Champaign, USA). The effect of topic variation in performance testing: the case of the chemistry TEACH test for international teaching assistants. *Language Testing* (London, UK), **16**, 1, 52–81.

Topic is believed to be an important test-method characteristic in many types of language tests. Some test developers choose neutral test topics in order to nullify the effect of background knowledge. Yet there is a need to determine how well prospective international teaching assistants (ITAs) can communicate within their own field. Recent studies have compared test results between general and field-specific oral English tests. However, field-specific performance tests for ITAs often provide different topics within the same field for each examinee, assuming equivalency between topics. The comparison of general topics versus field-specific topics is unable to capture the full effect of topic. This study reports research into topic features and the effect topic variation has on a particular performance test – the chemistry TEACH (Taped Evaluation of Assistants' Classroom Handling) test, designed for ITAs. Results suggest a relationship between topic of input (as defined by the topic features of concepts, maths and calculations) and test scores on the chemistry TEACH test.

99–635 Saville, Nick and Hargreaves, Peter (University of Cambridge Local Examinations Syndicate (UCLES), Cambridge, UK). Assessing speaking in the revised FCE. *ELT Journal* (Oxford, UK), **53**, 1 (1999), 42–51.

This paper describes the Speaking Test which forms part of the revised First Certificate of English (FCE) examination produced by UCLES and introduced for the first time in December 1996 [readers are referred to *First Certificate in English: Handbook*, UCLES, 1997]. The aim is to present the new test as the outcome of a rational process of test development, and to consider why the new design provides improvements in the assessment of speaking within the FCE context.

99–636 Schmitt, Norbert (U. of Nottingham, UK). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing* (London, UK), **16**, 2 (1999), 189–216.

In this paper the author argues that issues of construct validity should be given more prominence in the validation of lexical test items. One way of determining the construct validity of vocabulary items is to interview participants directly after taking the items to ascertain what is actually known about the target words in question. This approach was combined with the framework of lexical competency proposed by Nation (1990) in an exploratory study which investigated the behaviour of lexical items on TOEFL (Test of English as a Foreign Language). In individual interviews, six TOEFL vocabulary items were given to 30 pre-university international students who were then questioned about their knowledge of the target words' associations, grammatical properties, collocations and various meaning senses. The results suggest that the type of item currently employed in TOEFL does not adequately reflect association, grammatical and collocational knowledge, and that even meaning knowledge is not captured as well as might be hoped. This is taken to indicate that the field could benefit from deeper exploration of what vocabulary items are actually measuring.

99–637 Upshur, John (Concordia U., Montreal, Canada) **and Turner, Carolyn E.** (McGill U., Canada). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing* (London, UK), **16**, 1, 82–111.

Major differences exist in two approaches to the study of second language performance. Second language acquisition (SLA) research examines effects upon discourse, and is typically unconcerned with scores; language testing research investigates effects upon scores, generally without reference to discourse. Within a general framework of test-taking and scoring, the present authors report research from these two fields as it relates to questions of systematic effects on second language tests; and then examine findings incidental to a test-development project. The findings were consistent with language testing research into systematic effects of task and rater on ratings, and with SLA research into systematic effects of task on discourse. Using empirically derived scales as indicators of salient features of discourse, the authors infer that task type influences strategies for assessing language performance. Explanations for these joint findings are not afforded by either standard language testing nor SLA perspectives. There is no theory of method to explain how particular aspects of method affect discourse, how those discourse differences are then reflected in ratings and how task features influence the basis for judgement. It is concluded that a full account of performance testing requires a paradigm which incorporates relationships not specified in either the major language testing tradition or the tradition of SLA research.

Teacher education

99–638 Barbot, Marie-José (Université du Littoral, Côte d'Opale). Nécessité d'une formation

https://doi.org/10.1017/S0261444800014208 Published online by Cambridge University Press