PRACTICAL DESIGN AND IMPLEMENTATION CONSIDERATIONS OF A COMPUTER ADAPTIVE FOREIGN LANGUAGE TEST: THE MONASH/MELBOURNE FRENCH CAT

Jack Burston, Monash University and Monique Monville-Burston, The University of Melbourne

ABSTRACT

While the psychometric and statistical models underlying the design of computer adaptive tests (CAT) are well understood, relatively few working models exist for the purpose of foreign language assessment. Likewise, little has been published concerning the practical considerations affecting the implementation of such tests. In the process of constructing the Monash/Melbourne French CAT, we discovered much about putting testing theory into practice. The present paper reports this experience in three parts. In a preliminary section, we describe the academic context in which the French CAT was created and trialed. This is followed by a detailed consideration of the test presentation platform and operating algorithms. Lastly, we give an evaluation of the first administration of the French CAT, accompanied by a discussion of the test’s reliability and validity as a placement instrument for first year Australian university students.

KEY WORDS

Computer adaptive testing, item response theory, Rasch model, foreign language proficiency testing, French placement testing.

INTRODUCTION

Main frame computers have played an important role for some thirty years in educational measurement for item banking, statistical analysis, and automatic correction of standardized paper and pencil tests (Kaya-Carton, et al. 1991). In the last fifteen years, computers have been increasingly exploited to administer tests constructed in conformity with explicit psychometric hypotheses, most
notably those associated with item response theory (Lord 1980; Wright and Masters 1982; Wright and Stone 1979). During the past decade, the accessibility of powerful microcomputers has allowed the analytical potential of computers to be harnessed to improve testing procedures, resulting in a flurry of interest in the development of computer adaptive tests (Wainer 1983; Weiss 1982, 1983; Weiss and Kingsbury 1984). As a consequence, the psychometric and statistical models underlying the design of computer adaptive tests are now well understood (Wainer 1990). Most recently, attention in this field has focused on the evaluation of test results, i.e. their reliability and validity compared to written tests (Lunz and Bergstrom 1991) and empirical consequences of manipulating test algorithm parameters (Lunz, et al. 1993).

While computer adaptive tests (CAT) exist in a number of domains, few working models have been developed for the purpose of foreign language assessment. Likewise, relatively little has been published concerning the practical considerations affecting the implementation of such tests (Henning 1987; Larson 1991; Stevenson and Gross 1991; Tung 1986). In constructing our computer adaptive test for French, much of what we needed to know about putting testing theory into practice, of necessity, had to be learned first hand. The present paper reports this experience in three parts. In a preliminary section, we describe the academic context in which the French CAT was created and trialed. A detailed consideration of the test presentation platform and operating algorithms then follows. Lastly, we give an evaluation of the first administration of the French CAT, accompanied by a discussion of the test’s reliability and validity as a placement instrument for first year Australian university students.

BACKGROUND INFORMATION

Entry into First Year French at the University of Melbourne

Students coming to the University of Melbourne from the State of Victoria secondary schools and wishing to enroll in the first year French course must have passed the Victorian Certificate of Education (VCE) French examination. Successful completion of VCE French, however, is a rather poor indicator of the linguistic level of incoming students (Burston 1993). The latter manifest wide ability differences, from native/near-native speaking and writing skills to uneasy oral and written expression, with numerous lexical and grammatical errors. This lack of homogeneity is due to several factors: family background (Australian vs. French or francophone); increasing opportunities for in-country stays of various length, with or without formal instruction; type of school attended; type of course (correspondence or regular); and individual motivation and talent.
Placement levels

Previous experience had shown that for the first year course, it was convenient and useful to stream students. It had therefore been the practice, for the previous six years, to evaluate students upon their entrance and place them into three ability strata: advanced (A), standard (S), intermediate (I), representing respectively about 20%, 60% and 20% of student intake. The three levels compare approximately, to the ACTFL rating scale bands in the following manner: the I stratum corresponds to the Mid-Upper Intermediate range, the S stratum to the Upper Intermediate-Advanced range and the A stratum to the Advanced range (or higher). Students in their first year are expected to improve all their language skills and learn about aspects of contemporary French civilization; they are also introduced to textual studies. All lectures and class activities are conducted in French.

In order to assign students to one of the three groups, in recent years, the VCE French score was used, sometimes in association with a paper and pencil test — one of the French CAT predecessors — and/or other evaluation techniques (interviews, short compositions). For various reasons, particularly the hope to save time on test administration and grading, and the prospect of expeditious release of results, it was decided to develop a computer adaptive test.

THE TEST AND ITS PRESENTATION

Testing Model

The Monash/Melbourne French CAT is based on the Rasch model of Item Response Theory (Rasch 1960). IRT operates on the general assumption that it is possible to equate ability level (i.e. a unidimensional underlying latent trait) in a given domain directly with the difficulty level of specific tasks within that domain (Hambleton and Cook 1977; Traub and Wolfe 1981). The Rasch model of IRT goes about this by relating individual ability to a single parameter of test item difficulty on a common scale measured in units known as logits. The midpoint of this scale is set at 0.00, represented graphically on a vertical continuum. Harder items are located above this level and assigned increasingly higher positive values: e.g. + 1.00, + 2.00, + 3.00. In parallel fashion, easier items are located below the midpoint level and assigned increasingly negative values: e.g. -1.00, -2.00, -3.00. In accordance with an examinee's pattern of success or failure in responding to items of known difficulty, a continuously updated estimation is made of the probability of that person's answering a harder or easier question correctly. As the range of test item difficulty narrows to match a student's estimated ability, accuracy of prediction increases. Consequently, since a computer adaptive test may terminate as soon as an acceptable threshold of measurement error has been reached, only a fraction of potential test bank
questions actually needs to be presented to determine ability level with reasonable confidence. Aside from performing the required statistical calculations, the computer adaptive algorithm is also charged with selecting appropriate questions from the test bank inventory, i.e. those whose calibrated difficulty level most closely approximates the value targeted.

**Test Background**

The content of the **French CAT** derives from two paper and pencil diagnostic tests originally developed in 1975/1976. These focused on specific points of grammar across a comprehensive syllabus range. The first test identified areas of strength and weakness of incoming High School Certificate (HSC) students, i.e. final year secondary level entrants with a minimum of four years previous French, and was used to determine individual programs of remedial work. The second test assessed the language competence of students entering the advanced second year level stream. The combined question inventory, presented in a four item multiple choice format, totaled 125 partial sentence completions. In all, prior to their incorporation into the **French CAT**, the tests had been administered nearly 1200 times over a period of fourteen years.

In preparation for the computer adaptive test, the source archive of previously scored answer sheets was statistically analyzed to determine question difficulty levels and to ensure the fit of the data to the requirements of the Rasch model. The latter involves, in particular, eliminating from the item bank non-discriminating questions (e.g. those answered correctly or incorrectly by all examinees) as well as those to which observed responses were inconsistent with modeled expected responses (e.g. questions with a high difficulty value answered correctly by testees of low ability and vice versa). Since the progenitors of the **French CAT** had undergone similar statistical analyses as part of pilot testing undertaken during the early stages of their use, the Rasch infit statistics not surprisingly resulted in the rejection of only one item in the original pool of 125. The remaining survivors were assigned logit difficulty values varying from a high of + 3.50 to a low of -2. 10 (Tables 1A, 1B). Just under 100 test items were fairly evenly spread within a range of + 1.50 to -1.00, at an average measurement error below 0.13. In summary, both in terms of the quantity and quality of its question inventory, the **French CAT** database easily exceeded minimal requirements for the construction of a single parameter item response test (Henning 1984, 1991; Weiss and Kingsbury 1984).
| 3.5 .. | 3.2 .. | 2.7 .. | 2.3 .. | 2.1 .. | 2.0 .. | 1.8 .. | 1.7 .. | 1.6 .. | 1.5 .. | 1.4 .. | 1.3 .. | 1.2 .. | 1.1 .. | 1.0 .. | 0.9 .. | 0.8 .. | 0.7 .. | 0.6 .. | 0.5 .. | 0.4 .. | 0.3 .. | 0.2 .. | 0.1 .. | 0.0 .. | -0.1 .. | -0.2 .. | -0.3 .. | -0.4 .. | -0.5 .. | -0.6 .. | -0.7 .. | -0.8 .. | -0.9 .. | -1.0 .. | -1.1 .. | -1.2 .. | -1.3 .. | -1.5 .. | -1.6 .. | -1.7 .. | -2.1 .. |

Median

Table 1a. French CAT difficulty level calibration distribution. Average Standard Error Measurement: .128
Table 1b. Representative Test Questions/Difficulty Levels

Presentation Platform

The first practical considerations encountered in the actual development of the French CAT were those involving computer hardware and software. Since it was intended that the test be able to operate on the largest number of microcomputers, preference was given to the IBM-PC / MS-DOS environment. It was likewise decided to keep system requirements to a minimum: 384K RAM, single 360K floppy disk, any video adapter/monitor, no graphics, no mouse. Another important factor in the design of the test was the need to keep its programming structure modular and open-ended so as to easily accommodate database changes, i.e. revised question difficulty levels, addition of new test items, and future extension to other languages. Lastly, for obvious reasons of security, the contents of the run-time version of the French CAT needed to be encrypted. As an interim measure, a prototype was developed using HyperShell (Text Technology), a shareware authoring language. This worked so well that it was retained for the final version.
The primary requirement of the user interface for the test was that it be simple and resistant to mishap. In completing a short set of personal information input fields at the beginning of the French CAT, students learn all they need to know about the keyboard: i.e. location of essential keys, backspace deletion, etc. Since they are in a multiple choice format, answers to test items themselves require only entry of a single letter: A B C D. A practice section is nonetheless included to demonstrate operating procedures and allow a short warm-up period. In order to permit responses to be easily changed, they are not registered until they have been confirmed. As a counterbalance to hasty or accidental recourse to the <ENTER> key, confirmation is effected by pressing the <SPACEBAR>. Two special features of test items also necessitate particular explanation. Firstly, sentences may be complete as they stand, leaving the possibility of a null response. Secondly, because of the existence of discontinuous answers (e.g. ne ... pas,) double blanks may also occur. Contextualization of multiple choice alternatives is accomplished by displaying selected options in the sentence blanks. For example, given the stimulus je______ ai______ rien vu, typing the letter B to choose n’... ø results in that selection appearing in the sentence blank as: je n’ai rien vu (Tables 2a, 2b). Since tentative responses may be altered at will, testees are free to preview all selections in context before committing themselves to an answer.

THE DESIGN OF THE ABILITY ESTIMATE ALGORITHM

Ability Estimate Algorithm

The French CAT functions essentially like a language proficiency interview, presenting harder questions in reaction to correct answers and easier ones subsequent to incorrect replies. As with a human examiner, it is necessary to determine an appropriate difficulty level at which to begin the test. Theoretically, it is possible to accelerate the outcome of the test by setting its starting point higher or lower in anticipation of student competence (Tung 1986). In simulated test trials of the French CAT, however, fixing it elsewhere than at the 0.00 midpoint proved to have very little effect and could even be counter-productive. With mixed ability groups, which are much more the norm than the exception, beginning with harder or easier items would frequently require the administration of a greater number of questions to enable the algorithm to get back on the right track. Moreover, initiating computations too close to difficulty level extremes can cause the process to falter for want of suitable items.

Question Selection Algorithm

The choice of items for test presentation is determined by two considerations: the item difficulty level targeted and the item selection criterion used. In theory, a CAT may either adapt to target a specified pass/fall point or the current
estimated ability of the examinee (Lunz, et al. 1993). The latter method being the most effective with heterogeneous test groups, it was the one adopted for the French CAT. Item selection may be effected by one of two main procedures: Bayesian or Maximum Information (Weiss and Kingsbury 1984). Bayesian item selection is geared to minimize the expected posterior variance of the examinee's last ability estimate. As its name implies, the Maximum Information procedure seeks to select items that provide the maximum amount of item information at the last ability estimate. Although both methods tend to yield similar results, on relatively short tests Bayesian selection produces ability estimates which are regressed towards the prior estimate. Since the Maximum Information procedure
does not manifest this bias towards previous estimates, its use was given preference in the French CAT, the anticipated length of which was under 40 questions.

Although Maximum Information selection is free of the regressive bias of the Bayesian procedure, its use carries with it certain constraints. The most notable of these is the inability of Maximum Information selection to deal with cases of all right or all wrong responses. In order to establish upper and lower limits within which to calculate ability level, Maximum Information selection requires the elicitation of at least one right and one wrong response. This raises another practical test initiation consideration: that of the question selection procedure to follow pending the calculation of the first ability estimate. In the absence of the latter, i.e. from the second test item until the right/wrong condition is met, any increase or decrease in question difficulty level can only be determined heuristically. Setting a relatively large increment, e.g. ± 1.00 logit, can quickly cause the algorithm to run out of available test items. Setting a very small one, e.g. ± 0.10 logit, can lead to a prolonged succession of all right all wrong answers. On the basis of test simulations, the strategy adopted for the French CAT was to select an initial question set consisting of the five test items closest to the start-up level. These are then presented one after the other until either the right/wrong requirement is met or a sequence of all right/all wrong responses results. In the latter case, the algorithm advances in difficulty level increments of + 0.20 logits until an ability estimate can be made.

As should be apparent from the above discussion, one of the principal constraints on the design of a CAT is the size of its question bank. In developing the French CAT question selection procedure, it quickly became apparent that the minimal theoretical requirement of about "100 good quality items" (Tung 1986; Weiss and Kingsbury 1984) overlooks some very important practical considerations. Unless a question inventory numbers in the hundreds, just getting started can be problematic. Once the test is under way, uneven distribution of question difficulty levels is another source of complication. Whatever their provenance, test items will inevitably bear witness to the target group for which they were originally intended. As a consequence, difficulty levels will tend to cluster around one part of the spectrum, leaving greater or lesser gaps elsewhere. Given the advanced linguistic competence its progenitors aimed to assess, test items in the French CAT not surprisingly manifest an upper-end bias. The test median is in fact slightly above + 0.400, with over 80% of questions situated within + 1.00 logit of this.

Wherever patchy difficulty level distribution occurs, a problem of appropriate question availability is likely to arise. As items are used, they are automatically withdrawn from the test bank. This naturally creates ever greater gaps between the test items in thinner parts of the inventory. In seeking a more difficult question slightly above the + 1.00 level, for instance, it could happen that the
Presentation of this question could easily elicit an incorrect response, sending the algorithm on a futile quest for alternatingly easier/harder items. Lacking a suitable item at the upper estimated ability level, the test would be doomed to continue trying ever more difficult questions. The reverse situation, of course, could also occur by overshooting the mark in search of easier test items.

Obviously, the only way to avoid fruitless searches is to increase the size of the question bank. To be effective, however, additional items need to target the thinnest areas of the inventory, i.e. little would be gained by simply adding 100 items that end up in the well-represented sections of the difficulty range. Needless to say, setting out to create a few dozen questions around specific logit levels, e.g. -2.00, -1.75, +1.75, +2.00, is easier said than done. Though intuition, based on teaching experience, is an indispensable guide, item difficulty level can only be determined ex post facto through statistical analysis. New items are, of course, being written to fill in the French CAT question base. A reasonable estimate is that a total of about 200-250 items will be needed to achieve satisfactory results across a +2.50 logit difficulty range. In the interim, to prevent unproductive searches, question selection in the French CAT is monitored to ensure that the gap between the difficulty value sought and that chosen does not exceed +1.00 logit.

Stopping Conditions

A number of factors need to be taken into account in determining the point at which a computer adaptive test should terminate. The most critical of these is the margin of error expected of the ability estimate. Ideally, this should be as close to zero as possible. Since, however, standard error measurement (SEM) decreases in inverse proportion to the number of questions presented, the quest for high accuracy has to be balanced against the need to keep the test as short as possible. Reports in the literature regarding attainable accuracy vary considerably. One finds, for example, claims of a .105 SEM within 30 questions (Kaya-Carton, et al. 1991), .25 SEM within 20 questions (Bock and Mislevy 1982), .28 SEM within 50 questions (Lunz, et al. 1993). In the case of the French CAT, test simulations indicated that students could be reliably streamed into three broad ability groupings — high/mid/low — at a standard error below 0.35 within a maximum of 40 test items. In actual practice, virtually all tests which terminated at the .35 SEM threshold did so between 34-38 questions.

While the need to establish stopping criteria on the basis of standard error level and test length is well documented (Wainer 1990; Lunz, et al. 1993), other important considerations seem to have escaped notice. One such involves the maximum tolerable discrepancy between the item difficulty level sought for testing and the actual question chosen for presentation. As described above, in the absence of a uniform spread of logit values across the whole range of item difficulty levels, the test algorithm will at some point end up selecting questions
which get further and further off the mark on every cycle. The imposition of a maximum gap limit to prevent this constitutes in effect a stopping condition on the test. A similar pragmatic constraint on the continuation of a test may be encountered should the calculated ability estimate trigger a search for a question with a difficulty level beyond the upper/lower limits of the item bank. This is precisely the situation which results from setting the starting point of the test too close to its logit value extremes. Again, when this occurs, there is no practical alternative but to terminate the test. During the trials of the French CAT, the maximum gap limit terminated tests in about a quarter of the cases, on average after 24 questions at a SEM below .46. In comparison, less than 6% of the tests had to be abandoned due to the out-of-bounds condition.

TRIALING

First administration of the French CAT

The French CAT was administered to 114 students entering first year (non-beginners) French at the University of Melbourne at the beginning of the 1993 academic year. It was decided to trial the French CAT as a placement tool but to conjointly take into account VCE results. There were three reasons for adopting this procedure:

a) The correlation between the French CAT ability estimate values and performance relative to streaming levels was not yet known;

b) Since the French CAT is a test of grammatical knowledge, there was the potential to misjudge student abilities compared to the VCE, which is based on multi-skill assessment;

c) It offered the additional opportunity to compare French CAT scores and VCE results and, in the process, to assess the new VCE marking scheme, the validity and reliability of which had been called into question by many educators and secondary teachers.

The French CAT was administered during Orientation Week so as to allow streaming and enrollment in time for the beginning of classes. Five sessions were organized in a 24 seat capacity IBM compatible computer lab and were supervised by two staff members. As it turned out, one supervisor would have been sufficient. Staff gave a short presentation to introduce the test and guide students through the initial test directions and practice questions. No difficulty arose since the majority of students these days are computer literate and familiar with screens and keyboards. Testees found instructions simple and clear. Orientation required about ten minutes and students, working at their own pace, took between 15 and 35 minutes to complete the CAT.
Results and Follow up

Results were electronically collected and sorted according to decreasing ability estimate; these ranged from a high of +3.1 to a low of -2.58. Results are displayed in Table 3 (where the logits have been simplified for reading by retaining only one decimal and eliminating the decimal point).

The standard error measurement was generally 0.34 (the default stopping condition) with some discrepancies at the top and bottom of the ability range, e.g. 0.39, 0.50, and even 0.62 at the very top and 0.45, 0.63, 0.85 at the very bottom. This confirms, as indicated above, that the test is more discriminating in its center (roughly the S stratum) and less precise at its extremities. The average estimated ability logit value was +0.32. It is thus situated, for this group of testees, +0.3 logits above the criterion difficulty standard of the calibrated item bank scale (0.00) and about -0.1 logit below the test median (+0.40). (See above Question selection algorithm.)

The validity and reliability of the VCE results having been judged unsatisfactory, it seemed reasonable to place more confidence in the ability measures provided by the CAT (Burston 1993). The French CAT was therefore adopted as the basic streaming instrument. For practical reasons dictated by class size considerations, it was decided to set the cutoff logit value at +1.08 for the A group and -0.49 for the I group. A high or low score on the VCE/HSC was taken as confirmation criteria for this placement. In the case of wide discrepancy between ability estimated by the CAT and that measured by the VCE/HSC, students were...
individually interviewed and given a short essay to write for assessment. A few further adjustments were made during the first two weeks of classes after tutors had had the opportunity to observe their students in the classroom. Unquestionably, the French CAT measurements allowed much more accurate placement than the VCE/HSC results would have provided.

Discussion

Reliability and Validity

The paper and pencil tests from which the French CAT originated were designed as achievement tests. They were based on the Year 12 (final year of secondary studies) syllabus which explicitly specified which areas of French grammar should be mastered for the HSC. The two tests were used at university level to check grammatical competence (see above Test background).

Passing from traditional test to CAT involved, along with a change in medium, a change in purpose, i.e. a shift from achievement to placement/proficiency testing. When the CAT was trialed, therefore, two essential questions to be addressed were that of its reliability and that of its validity. Reliability of the CAT was not a major concern, inasmuch as the paper and pencil tests from which it derived had regularly proved to give consistent results. Several studies (Weiss 1982; Weiss and Kingsbury 1984; Lunz, et al. 1993) have shown that the reliability of computer adaptive tests does not vary significantly from that of their written test counterparts. It could be expected, therefore, that results obtained with conventional vs. adaptive formats would yield similar distribution of testees.

As with reliabilities, validities too remain comparable in transfers from traditional tests to CATS. The content validity of the French CAT was thus not at issue, since the new test contained about the same balanced sample of grammatical items as the original tests. Above all, then, what remained to be appraised was the empirical validity of the French CAT as a measurement of the general linguistic competence of university entrants and its potential use for placement purposes. Since, according to IRT, the ability to be measured is presupposed to be unidimensional (in the present case, French morphosyntax,) there was the potential that this trait might prove to be too restrictive. In principle, given that other skills were not evaluated or were evaluated only in a limited way, we could not assume that the CAT would be an appropriate placement instrument. In actual practice, as the discussion below demonstrates, the French CAT has proven to be a satisfactory tool for sorting first year students into the three intended strata. After several weeks of teaching and observation, tutors confirmed that, with a few exceptions (see below Misclassifications) students had been properly categorized.
Misclassifications

For reasons already mentioned, it was decided to ignore the VCE results for placement. Incoming students were assigned to the three strata according to their score on the French CAT. The usual distribution for first year streaming (A: 20%; S: 60%; I: 20%) was retained. This resulted in creating cutoff points at logit + 1.10 [A/S] and logit -0.50 [S/I]. After several weeks of monitoring students' performance in all skills, tutors reported that a total of thirteen students may have been misplaced by the French CAT (See Table 3):

a) Five who were placed in the Advanced stream were deemed overrated. One case only (+1.4 logit,) however, represents a clear misclassification. Instructors hesitated in their assessment of the +1.7 and +2.0 logit level students. The two others, -with ability estimates of +1.1, are borderline at the bottom of the A band. It 'would have been more natural, given configuration of the distribution curve, to set the cutoff point at +1.2; but the latter was lowered to better balance class sizes.

b) Three students (-with abilities 0.0, 0.0 and +0.5) were also thought to have been over-placed in the S stream and would have been better placed in the I group.

c) The French CAT underrated two students in the S stream who, despite +0.8 and +0.9 logit estimates, were judged by their tutors to be have advanced level proficiency.

d) Of the three I students which the French CAT underestimated, instructors expressed hesitation in the case of one (-0.9) and one was borderline (-0.6, i.e. -0.1 logit below the cutoff point). The third one (-1.9, with standard error of measurement 0.72) was obviously a misclassification.

In total then, only seven students were unequivocally misplaced by the French CAT, i.e. 6.% of the testees. Given the extraneous affective factors which inevitably accompany testing of any type during Orientation Week (personal circumstances, stress, total bewilderment, etc.,) this represents a quite satisfactory error in placement rate.

Further Theoretical Considerations: the Question of Construct Validity

For some years, testing and language teaching specialists have raised the issue of the construct validity of assessments based on a single measurement trait (e.g. reading ability, grammatical competence, etc.,) as is the case in IRT (Henning, et al. 1985; Canale 1986; Tung 1986; Hamp-Lyons 1989; Lange 1990; Madsen 1991; Meunier 1994; etc.). Criticism is voiced on the grounds of the multifaceted mature of language mastery: "How feasible is it to test a multidimensional
Construct with an unidimensional approach such as IRT?" (Meunier 1994, 28). Though understandable, such objections do not fully appreciate the nature of unidimensionality in the context of IRT. As McNamara (1991) makes clear, concerns about the construct validity of unidimensionality are marked by a failure to distinguish two types of model: a measurement model and a model of the various skills and abilities potentially underlying test performance. These are not at all the same thing. The measurement model posited and tested by IRT analysis deals with the question "Does it make sense in measurement terms to sum scores on different parts of the test? Can all items be summed up meaningfully? Are all candidates being measured in the same terms?" This is the "unidimensionality" assumption. [... I As for the interpretation of test scores, this must be done in the light of our best understanding of the nature of language abilities [... ].(172-3).

IRT makes no direct pronouncements about the different linguistic skills that test results may presuppose. It simply assumes a single dimension relating ability and difficulty. From this point of view, what has been said above (Section 1) about the construction of the French CAT insures the uniformity of its measurements.

Another validity issue which could be raised is that the French CAT is intended to be the basic first year placement instrument (-with complementary ones, such as short interviews or a listening comprehension test) and that it is grammar-based. This touches more upon the traditional notion of test face validity of - i.e. its public acceptability in a language teaching world where communicative approaches prevail. Undeniably, overall language competence is the result of various components: conversational competence, reading competence, grammatical competence, sociolinguistic and cultural awareness, etc. This is certainly reflected in the contents and teaching methodology of the first year language course at the University of Melbourne (cf. end note 4).

Notwithstanding, given that between a third and a half of our first year intake goes on to complete a three year major" in French, accuracy is highly prized across all skills in the curriculum. In this pedagogical context, at least, the analytical nature of the French CAT and its morphosyntactic focus poses no problem of face validity.

Taking grammatical accuracy as a yardstick for placement in first year university level courses has proved effective. There is a high correspondence between the grammatical competency demonstrated by our first year students and their proficiency in other language skill areas (Burston, Monville-Burston, and Harfouch forthcoming). The French CAT has also shown good short-term predictive validity (Burston, Monville-Burston, and Harfouch 1995). This is not particularly surprising, since knowledge of the structure of a language (as knowledge of its lexicon) pervades all basic skills, Our findings are confirmed, furthermore, by various studies which demonstrate that grammatical
competence consistently correlates with other linguistic abilities (Oller 1976; Farhady 1983; Brown and Bailey 1984; Carlson and Camp 1985; McNamara 1990). Moreover, it is worth noting that with discrete point grammar testing washback effects are avoided, now that teaching methods in secondary schools focus on developing communicative skills rattler than the production of correct forms. At least as long as the teaching of grammar remains out of favor, there is little danger of examinees being trained to take and excel in grammar tests.

CONCLUSION: THE FUTURE OF FRENCH CAT

Evidence has been provided that the French CAT is reliable and that it has content, construct, empirical and, for those not adverse to the teaching of grammar, face validity. Practical evaluations by experienced teachers have validated placement decisions based on ability estimates made by the test. The French CAT has fulfilled its purpose: it is an efficient instrument which can be used with confidence to quickly stream students entering first year French programs in Australian universities, provided of course that the limitations of the simple measure it gives are understood and the results intelligently interpreted (Bachman 1990; Davies 1992; Spolsky 1995).

Several further developments - practical as well as research oriented - are planned in order to improve and enrich the test, and to understand better how it operates. A protocol study has been undertaken to investigate the criteria and strategies used by students in choosing answers to questions presented in the French CAT (Warren, forthcoming). The information gathered will be useful as new items are added to the pool, for it is planned to extend the upper and lower ranges of the test to assess a wider ability range than is currently possible. The creation of a larger question inventory will, furthermore, serve to bridge gaps between the difficulty levels of items, which in turn will help to reduce the standard error of measurement. Such improvements will be all the more if, as expected, the French CAT is called upon to evaluate a larger, even more heterogeneous student population. With consideration presently being given in several Australian universities to disassociating language acquisition from year level, it will be important to define degrees of proficiency, and be able to quickly and appropriately assign students to ability groups. For this purpose, five groups (with three cutoff points) — as compared to the present three (with two cuts) — will be needed. The feasibility of sorting testees into a larger number of strata will, thus, need to be studied. Finally, it is planned to develop a second, richer, adaptive test incorporating listening comprehension based on cloze dictation. This will be another context for exploring further the advantages and problems of computerized adaptive testing experienced with the French CAT.
NOTES

1 The best known foreign language CATs are those developed in 1987 at Brigham Young University for French, German, Spanish, and ESL (F-CAPE, G-CAPE, S-CAPE, CALI). A French reading proficiency CAT was produced in 1988 for the American Council on the Teaching of Foreign Languages. In 1990 an ESL CAT (ToPE) was released by the University of Edinburgh.

2 This new end-of-secondary studies certificate replaced the High School Certificate (HC) in 1992.

3 Very advanced students may be asked to directly enroll in second year French. If, because of a timetable clash for example, students have difficulties enrolling at the appropriate ability level (I, S, or A), they may obtain permission from the course coordinator to move to a contiguous level. Students, however, are strongly advised to remain, for one semester, in the stream in which they have been placed. Upon end of semester assessment, if they have underachieved or overachieved, they may be invited to change streams. This procedure permits flexibility and aims at having students learn at their particular ability level.

4 Students admitted to the Intermediate stream use textbooks such as B.G. Hirsch and C.P. Thompson, Ensuite or the Nouveau Sans Frontières 2 (CLE International). At the Standard level, typical teaching materials include C. Carlut and W. Meiden, French for Oral and Written Review; N. Mauchamp, La France d’aujourd’hui; A. Camus, L’Exil et le royaume; A. Ernaux, La Place, Prévert, Paroles. The same applies to the Advanced level, but the program is more intensive and enhanced by more complex tasks.

5 Since trialing began, a Macintosh version of the French CAT has also been developed.

6 A prepackaged CAT shell, the MicroCAT Testing System (Assessment Systems Corporation) was also considered, but not pursued due to its developer’s lack of interest in providing test details.

7 Variants of the Bayesian procedure also exist (e.g. EAP, MAP), but differences between them are not relevant here.

8 During “Orientation week,” which immediately precedes the teaching period, first year students are invited to participate in a variety of activities which familiarize them with their new educational environment and prepare them for their course.

9 For similar concerns in relation to reading comprehension, see Canale 1986.

10 Inadequacy of placement is determined by the lack of correlation between CAT and VCE/HSC results and, more crucially, by tutors’ perceptions. Doubtful cases (indicated by a question mark on Table 3) represent students who proved difficult to judge for various reasons: irregular attendance, irregular quality of assessed exercises, effect of outside assistance from private tutors, etc. In 1994, the French CAT was trialed again, but this time in conjunction with other testing instruments (a self-assessment test, a short interview and a listening comprehension test), to evaluate further its concurrent validity—and thus consolidate its face validity. The results are reported in Burston, Monville-Burston, and Harfouch (forthcoming). The predictive validity of the French CAT is discussed in Burston, Monville-Burston, and Harfouch 1995.

11 A Japanese CAT based on the French prototype has now also been developed in collaboration with the Language Research Testing Centre at the University of Melbourne.
REFERENCES


AUTHOR'S BIODATA

Jack Burston (B.A./M.A. in French, Ph.D. in Linguistics) is Senior Lecturer in French, Department of Romance Languages, Monash University. Monique Monville-Burston (M.A./Ph.D. in Linguistics) is Associate Professor in French, Department of French and Italian Studies, The University of Melbourne. The authors are presently completing a comprehensive CALL grammar French review program for use with intermediate/advanced level university students. Jack Burston teaches a post-graduate course in CALL and supervises M.A. and Ph.D. candidates in
this area. He is particularly interested in authoring languages and systems for courseware
development and is Chair of the CALICO Hypermedia SIG.

The authors would like to hear from other colleagues involved in the development of computer
adaptive language tests. A demonstration version of the French CAT is available and may be
obtained on request.

**AUTHORS' ADDRESSES**

Jack Burston
Department of Romance Languages
Monash University
Clayton, Victoria
3168 Australia

Phone: + 61-3 9905-2222
Fax: +61-3 9905-2137
E-mail: frn373b@alpha1.cc.monash.edu.au

Monique Monville-Burston
Department of French and Italian Studies
The University of Melbourne
Parkville, Victoria
3052 Australia

Phone: +61-3 9344-5168
Fax: +61-3 9347-2489
E-mail: monique_burston@muwayf.unimelb.edu.au