Graduate
Management
Admission
Council®

Development of a Multiple-Component CAT for Measuring Foreign Language Proficiency (SIMTEST)

Mick Sumbling, Pablo Sanz, M. Carme Viladrich, Eduardo Doval and Laura Riera

**Universitat Autònoma de Barcelona
Catalonia, Spain**

Presented at the June 8, 2007 Poster Session

Abstract

This paper describes the test development of a multi-component CAT called SIMTEST, which has been designed to measure foreign language proficiency in terms of the Common European Framework of Reference (CEFR) for languages. The background to test development is described in its institutional and historical settings, i.e. in the context of a Spanish university language service working initially with other universities in Catalunya in order to develop a common inter-university examination of students' foreign language skills. SIMTEST is presented in terms of its component parts and the six CEFR levels into which it classifies examinees. Item bank development details are provided, including item pre-testing administration methods, number of participants, calibration criteria, organization of items and anchors in pre-testing packs, the total of items pre-tested and the number banked for operational use. One-parameter IRT difficulty distributions had a wide range of difficulties for the 775 items currently banked, with the majority at an intermediate level of difficulty. The majority of Pearson correlation coefficients between certification components and their standard deviations were approximately 0.70. The pros and cons of SIMTEST are described with reference to various surveys carried out with teachers and students. The paper concludes with plans for future modifications, which include the incorporation of an automatic pre-testing element within the operative test, adjustment of the algorithm to allow for more balanced content distribution and a commitment to improve the quality of communication between all stakeholders—examinees, teachers, and testers.

Acknowledgments

Presentation of this paper at the 2007 Conference on Computerized Adaptive Testing was supported in part with funds from GMAC®.
Thanks and acknowledgements are also due to Rebeca García-Rueda

Citation

Sumbling, M., Sanz, P., Viladrich, M. C., Doval, E., & Riera, L. (2007). Development of a multiple-component CAT for measuring foreign language proficiency (SIMTEST). In D. J. Weiss (Ed.). ***Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.*** **Retrieved [date] from** www.psych.umn.edu/psylabs/CATCentral/

Author Contact

Mick Sumbling, Language Testing Unit (UAC), Servei de Llengües, Edifici M1. Campus de la Universitat Aútonoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.
E-mail: sim.crac@uab.cat

Development of a Multiple-Component CAT for Measuring Foreign Language Proficiency (SIMTEST)

Background to the CAT

SIMTEST, the computerized adaptive test (CAT) described below, was designed and developed at the language service (Servei de Llengües) of the Universitat Autònoma de Barcelona (UAB) in Catalonia, Spain. A language service is a typical feature of Spanish universities; providing as it does a school of languages, translation service, and other linguistic resources primarily, but not exclusively, to members of the university community. At the time of the origins of the SIMTEST, the language service at the UAB was known as SIM (Servei d'Idiomes Moderns).

From 1998 until the summer of 2000, SIM was involved in an interuniversity exam project with the language services of three other Catalan universities in response to a perceived mutual need to provide common local examinations of EFL (English as a Foreign Language) and to unify evaluation criteria at the four centers. This collaboration, the initiative for which had come from the centers themselves, led to exams that were essentially traditional in both their format (paper-and-pencil) and their content (e.g., "rational" cloze tests).

When this first interuniversity exam project ended in September of 2000, due largely to political reasons, a decision was made at SIM and a brief given to continue our research in testing and to design and develop a more innovative, computer-based test that would detect the common reference levels of language proficiency referred to and described in the document "A Common European Framework of Reference (CEFR) for Languages: Learning, Teaching, Assessment" (Council of Europe, 2001). At this stage, the document, which contains a collection of calibrated descriptors of language proficiency at the core of its system of levels, was the 1998 draft version. The brief also put an emphasis on initially designing an English placement test, the development of which might also lead to versions detecting level of proficiency in other languages.

Research and trialing led to the design and initial development of a computer-based test which quickly caught the attention of the Catalan Autonomous Government (*Generalitat de Catalunya*). At the initiative of the Ministry of Universities, Research and Information Technology, a second interuniversity exam project was begun in September 2001. This time, the project involved the language services of all the Catalan universities in response to a draft government bill requiring students to graduate with a minimum level (CEFR level B2.1) in a third language, i.e. a language other than Catalan and Spanish. The language centers of the eight major universities (six public, one private and one virtual), who were directly involved, now agreed to use SIMTEST as the central element of a system which would detect, evaluate and certificate graduates' language proficiency in a third language (known as PUC, i.e. *Prova Universitària de Competència*).

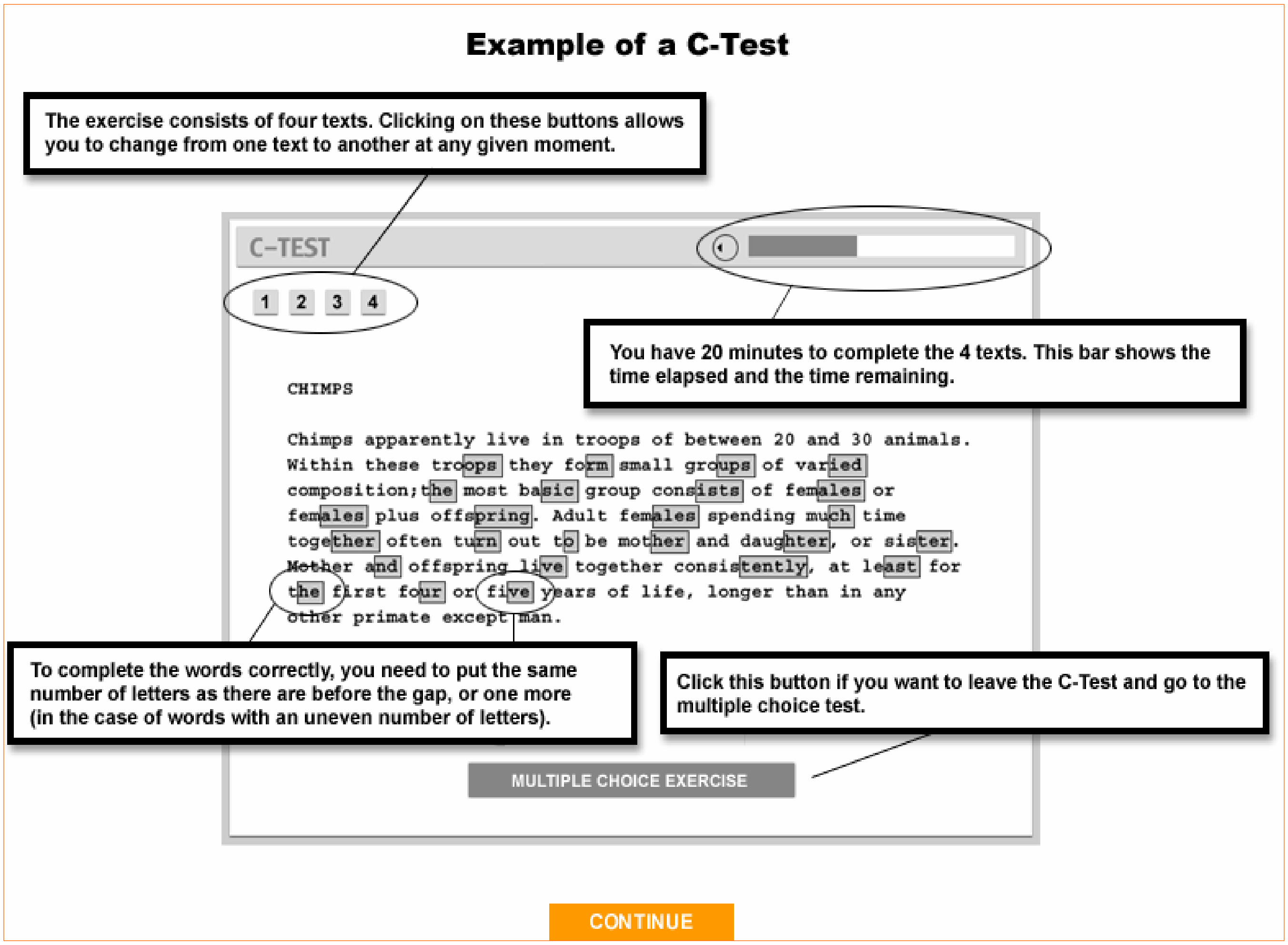
The collaboration lasted for two years. Then, a change of Government (and language policy) put an end to the PUC in November 2004. The two years of regular weekly meetings between representatives of the different centers had provided a think-tank which could explore the possibilities of using a CAT in conjunction with other language tests to the end of providing a profile of ability in different language skills. Valuable discussions had taken place on how to apply and standardize evaluation criteria, train raters, and so on. There was also the opportunity to pilot CAT procedures and pre-test items with larger numbers of students than would otherwise have been possible. All of this would contribute greatly to the development of SIMTEST 2.0, the current operational version for local and on-line use and in both placement and certification contexts.

Description of the Test

SIMTEST 2.0 classifies examinees according to six levels of proficiency as defined by the Council of Europe (2001). It consists of four component tests (C-test, VGF-CAT, Listening – CAT and Self-assessment CAT) that may be used in different combinations, according to the testing context or institutional requirements.

The C-test (see Figures 1 and 2) is the only non-CAT component of SIMTEST and is “based on the same theory of closure or reduced redundancy as the cloze test” (Alderson, 2000). C-tests were developed in Duisburg, Germany in the 1980s as a response to cloze testing. In SIMTEST the candidate completes four C-tests, each of which contain 25 words lacking the second half of their letters. Results are linked to a scale indicating the CEFR level. Serving as an initial indicator of the candidate’s global level, they also provide an entry point to the following CAT.

Figure 1. C-Test Instructions (SIMTEST Local Version)



Example of a C-Test

The exercise consists of four texts. Clicking on these buttons allows you to change from one text to another at any given moment.

You have 20 minutes to complete the 4 texts. This bar shows the time elapsed and the time remaining.

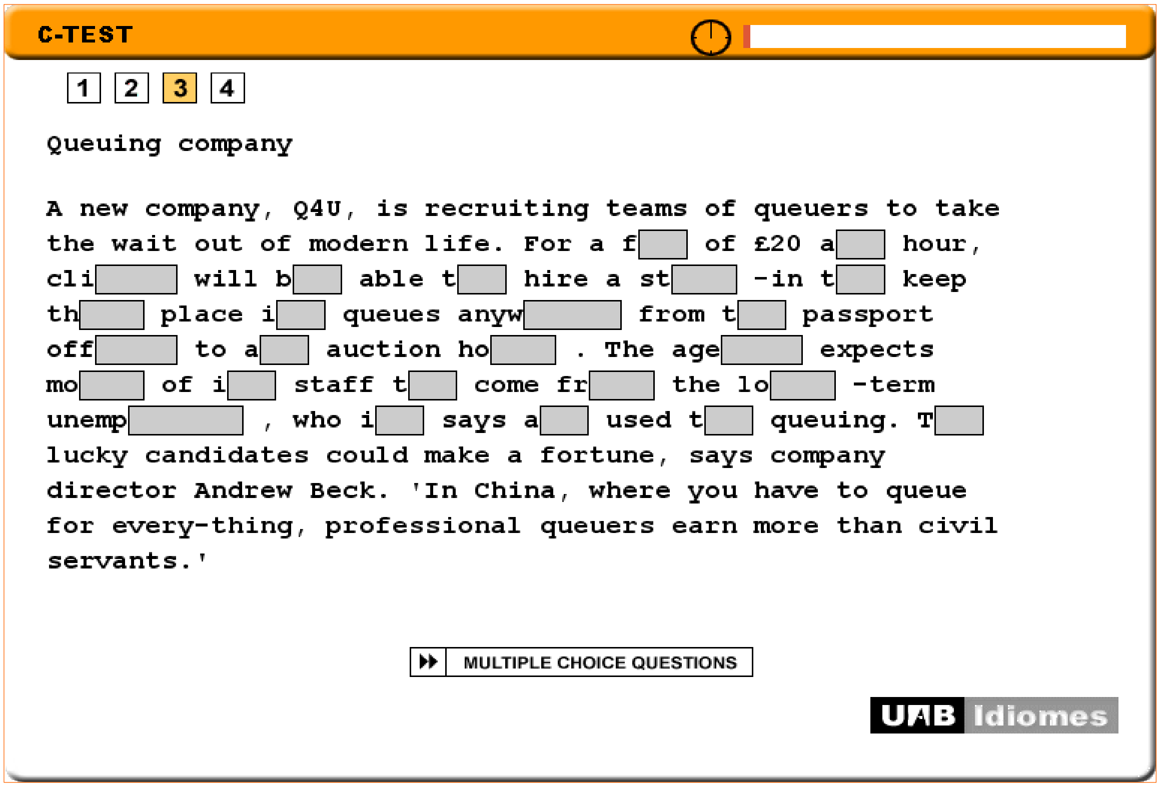
CHIMPS

Chimps apparently live in troops of between 20 and 30 animals. Within these troops they form small groups of varied composition;the most basic group consists of females or females plus offspring. Adult females spending much time together often turn out to be mother and daughter, or sister. Mother and offspring live together consistently, at least for the first four or five years of life, longer than in any other primate except man.

To complete the words correctly, you need to put the same number of letters as there are before the gap, or one more (in the case of words with an uneven number of letters).

Click this button if you want to leave the C-Test and go to the multiple choice test.

Figure 2. Example C-Test (SIMTEST Local Version)

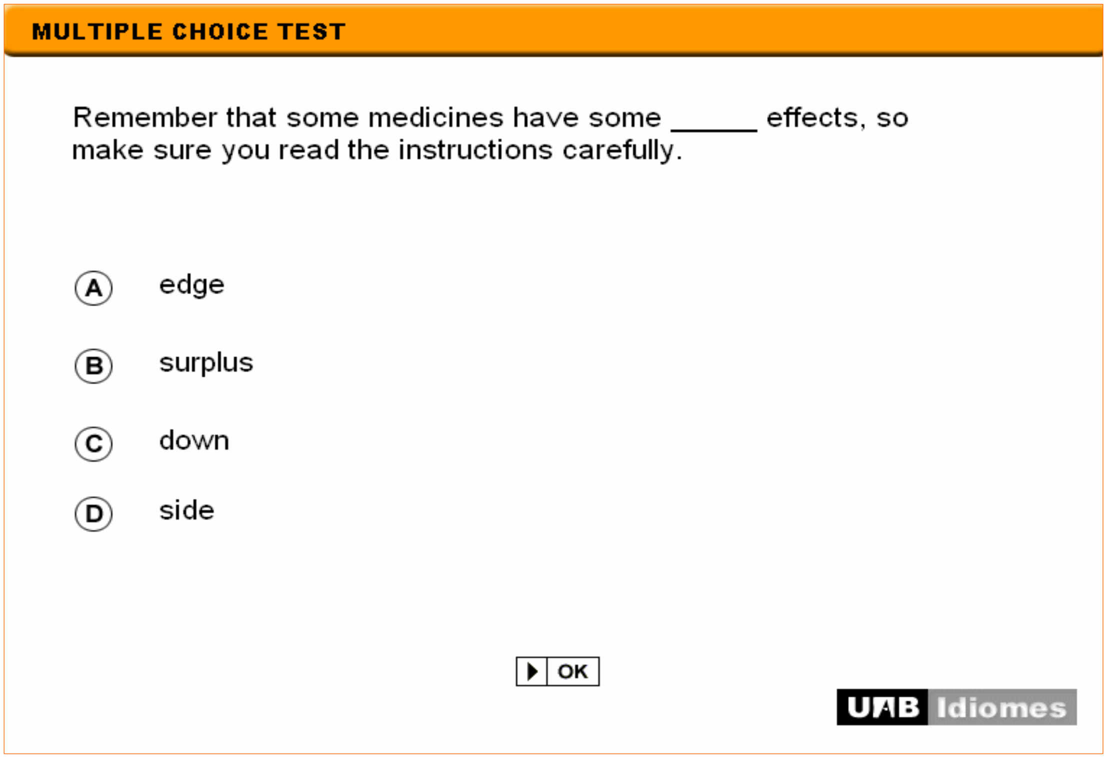


Queuing company

A new company, Q4U, is recruiting teams of queuers to take
the wait out of modern life. For a f___ of £20 a___ hour,
cli___ will b___ able t___ hire a st___ -in t___ keep
th___ place i___ queues anyw___ from t___ passport
off___ to a___ auction ho___ . The age___ expects
mo___ of i___ staff t___ come fr___ the lo___ -term
unemp___ , who i___ says a___ used t___ queuing. T___
lucky candidates could make a fortune, says company
director Andrew Beck. 'In China, where you have to queue
for every-thing, professional queuers earn more than civil
servants.'

In both placement and certification situations, the C-test provides the entry point to the CAT testing knowledge of vocabulary, grammar, and communicative functions, known as the VGF-CAT (see Figure 3 for an example). A series of multiple-choice questions appear one at a time on screen in an adaptive administration.

Figure 3. Example of a VGF-CAT Item (SIMTEST Local Version)



MULTIPLE CHOICE TEST

Remember that some medicines have some ______ effects, so make sure you read the instructions carefully.

Ⓐ edge

Ⓑ surplus

Ⓒ down

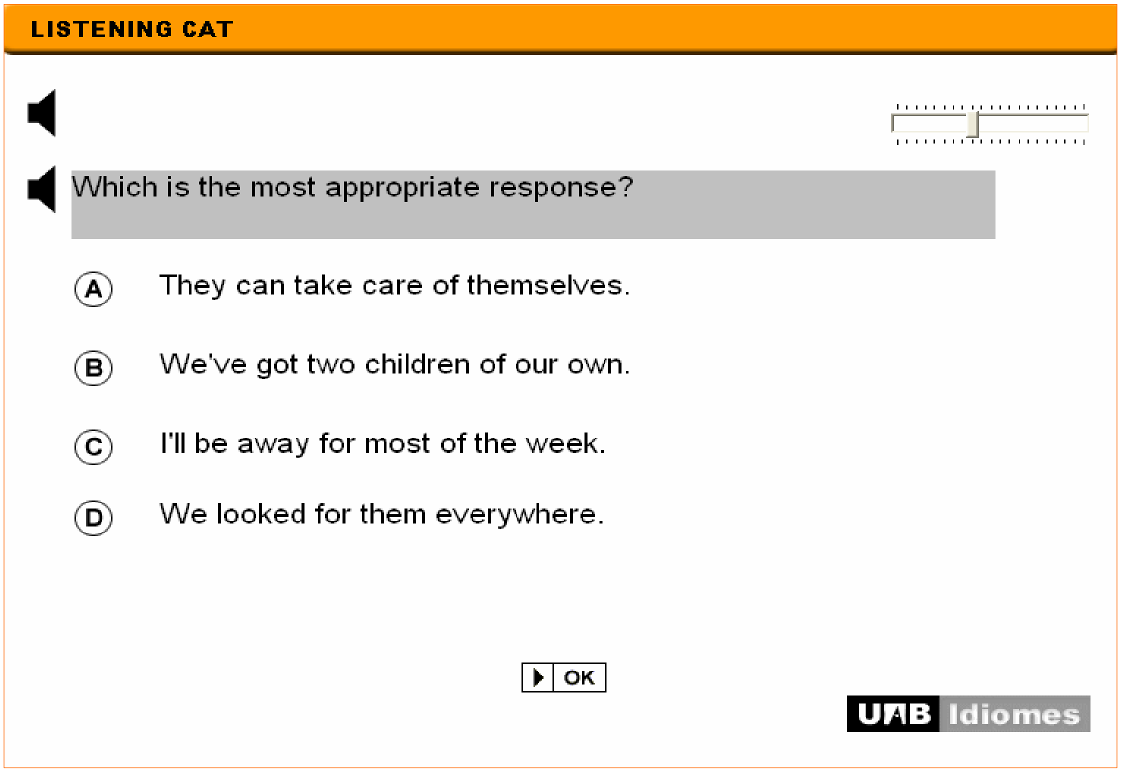
Ⓓ side

In certification situations the Listening-CAT follows the VGF-CAT. This test consists of series of multiple-choice questions testing listening comprehension of gist and details, recognizing situations, or identifying an appropriate response to an utterance (see Figure 4 for an example of a Listening-CAT item).

A self-assessment-CAT, **i**ntended for first-year undergraduates and presented to them in CD-ROM format, tests candidates' perceived ability to perform tasks in the four language skills (reading, writing, listening, speaking) by reference to "can-do statements" (Figure 6) about ability in the four language skills adapted from the CEFR.

The English placement test currently in use at the *Servei de Llengües* consists of C-tests and VGF-CAT items from the placement item pool used in conjunction with a (20 minute) writing sample and a brief (5 minute) individual interview. For certification purposes, C-tests, VGF-CAT, and Listening-CAT items from the certification pool are combined with a 90-minute writing paper and a paired (20-minute) oral exam. The self-assessment test is currently under review for use on-line, having previously only been available to fresher students in CD-ROM format.

Figure 4. Example of a Listening-CAT Item



LISTENING CAT

Which is the most appropriate response?

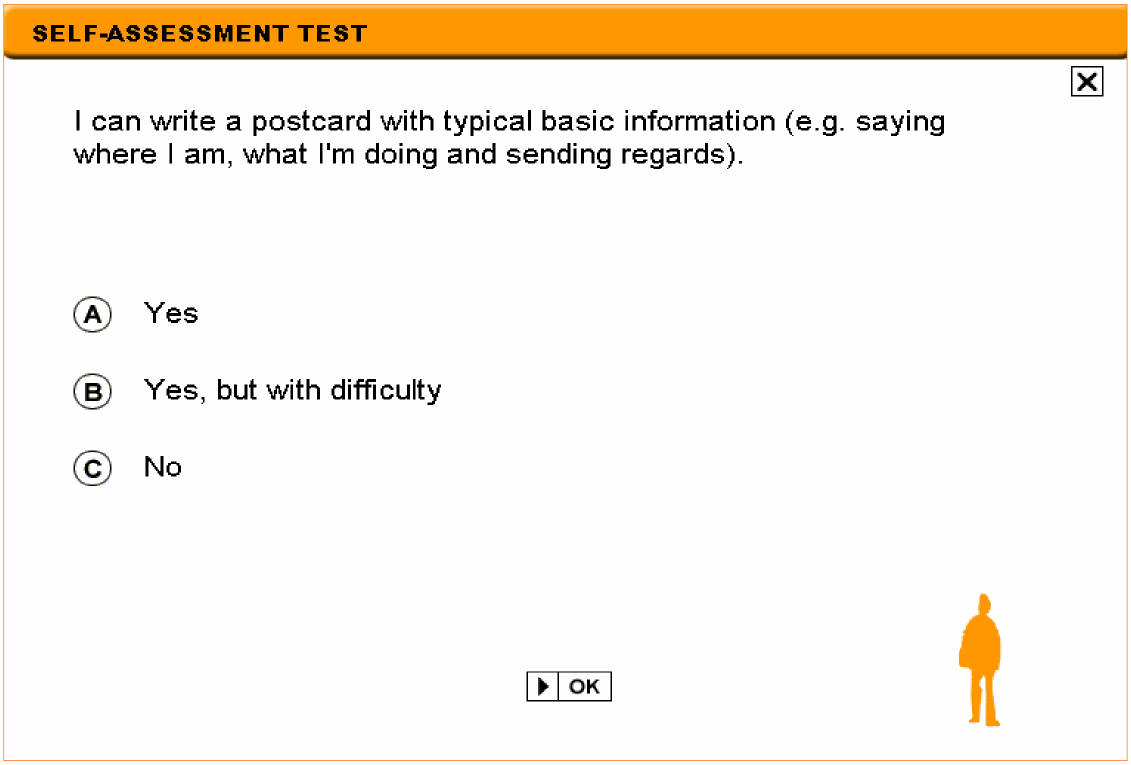
Ⓐ They can take care of themselves.

Ⓑ We've got two children of our own.

Ⓒ I'll be away for most of the week.

Ⓓ We looked for them everywhere.

Figure 5. Example of a Can-Do Statement (CD-ROM Version)



SELF-ASSESSMENT TEST

I can write a postcard with typical basic information (e.g. saying where I am, what I'm doing and sending regards).

Ⓐ Yes

Ⓑ Yes, but with difficulty

Ⓒ No

Algorithm for the CAT Components of the Test

An adaptation of the statistical item selection algorithm, mentioned in Dunkel (1999) and based on an algorithm described by Henning (1987) works as follows:

1. The examinee responds to an item at a level determined by the C-test result or, if the C-test is not taken, at a random level.
2. The program chooses the next three appropriate items. Each item will be one level of difficulty above the previous item if the examinee answered that item successfully, and one level below if the examinee answered that item incorrectly.
3. Once four items have been encountered, the program routinely estimates ability (θ) and associated error of estimate (s) after each item encountered. To do this, use is made of the following approximation formulas:

$$\theta = h + w(f - 0.5) + \ln\left(\frac{1 - \exp(-wf)}{1 - \exp[-w(w - wf)]}\right) \quad (1)$$

$$s = \left(\frac{w}{L}\right)\left(\frac{1 - \exp(-w)}{(1 - \exp(-wf))(1 - \exp[-w(w - wf)])}\right)\frac{1}{2} \quad (2)$$

where h = test height of mean difficulty of item encountered at each point, considered cumulatively [the sum of the item difficulty estimates (*d)* divided by the number of items encountered (L)]:

$$h = (\sum d\,)\,/\,L \quad (3)$$

w = test width of the span of item difficulties encountered, represented as the following quantity [this formula averages the two highest ($d2$) and lowest difficulties ($d1$) encountered in deriving test width, in order to provide greater accuracy]:

$$w = [(dL + dL - 1 - d2\ - d1)/2] \times [L\,/(L - 2)] \quad (4)$$

f = proportion of correct responses and r = number of correct responses:

$$f = r/L \quad (5)$$

With SIMTEST, a standard error of estimate of 0.5 can be achieved with as few as 6 items encountered, although on average 15 items are required. If this level of accuracy is not attained within 30 items, which has never happened in the operational phase, the program can be terminated on the grounds that the respondent is misfitting the measurement model by responding arbitrarily (Henning, 1987).

For the self-assessment CAT the algorithm functions slightly differently from the description above. With the 'Yes, but with difficulties" option the algorithm varies the difficulty of the next item, so that in two out of every three times the next item represents a higher level of difficulty.

Item Bank Development for the VGF-CAT Component

After initial research had pointed quite clearly to the use of a CAT administering multiple-choice items focusing on knowledge of vocabulary and structure, item sources—including the existing 100-item placement test and a large pool of items originally designed for language practice—

were identified. From the more than 3,000 items available, a pre-selection of 500 was made and these items were then vetted, where necessary edited, and prepared for pre-testing.

Results from the first pre-testing sessions in May 2001 were encouraging (see Table 1). Three hundred and sixty-two (72%) of the 500 items pre-tested were shown to be psychometrically suitable for inclusion on a CAT administration. There was also a good spread of item difficulties and discrimination across the six course levels. This initial analysis was based on classical test theory (CTT).

In September 2001, the prototype CAT, now known as SIMTEST 1.0 was piloted in placement sessions at the center. Validation studies made at the time gave encouraging results in the form of correlations with the other components of the placement test battery. Pearson correlation coefficients were 0.71 for oral exam and final placement scores, and 0. 69 for writing sample score.

Table 1. VGF Item Bank Development 2001 to 2004 (See Text for Details)

Year:	2001	2003	2004	2005
Item administration method:	Paper & pencil	Computer-based	Computer-based	Computer-based
Participants:	225	405	324	396
Calibration criteria:	CTT	CTT	IRT	IRT
Organization:	5 packs x 100 item	5 packs x 68/70 items 22 anchor	5 packs x 80/81 items 12 anchor	5 packs x 40 items 10 anchor
Total items pre-tested:	500 (new)	300 (new)	362(old)+6(new)	10(old)+150(new)
Banked items accumulated:	362	629	625	775

Further computer-based pre-testing sessions took place in 2003, 2004 and 2005. (see Table 1). The design included anchor items between packs and across years, and from 2004 item analysis was based on the 1-parameter Rasch model (1PM). Efforts were made to increment the ratio of examinees per item. Results presented below include the Rasch calibration of the 775 items banked during this period.

The 1PM IRT difficulty distributions for the 775 items in the item bank (Figure 6) show a wide range of difficulties represented with the majority of items at an intermediate level difficulty.

The box-plot in Figure 7 shows the proportion of correct responses according to the course level of students participating in pre-testing. As might be expected, the higher the level, the higher the proportion of correct responses.

Figure 6. 1PM IRT Difficulty Distributions for the 775 Items in the Item Bank

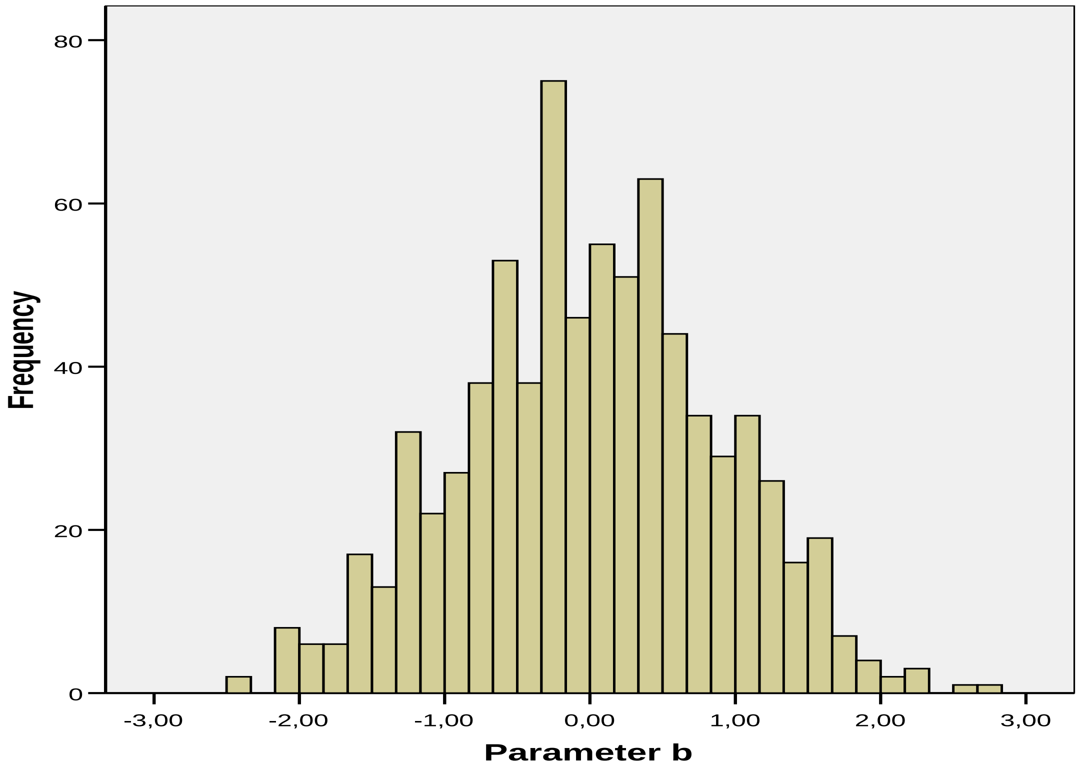
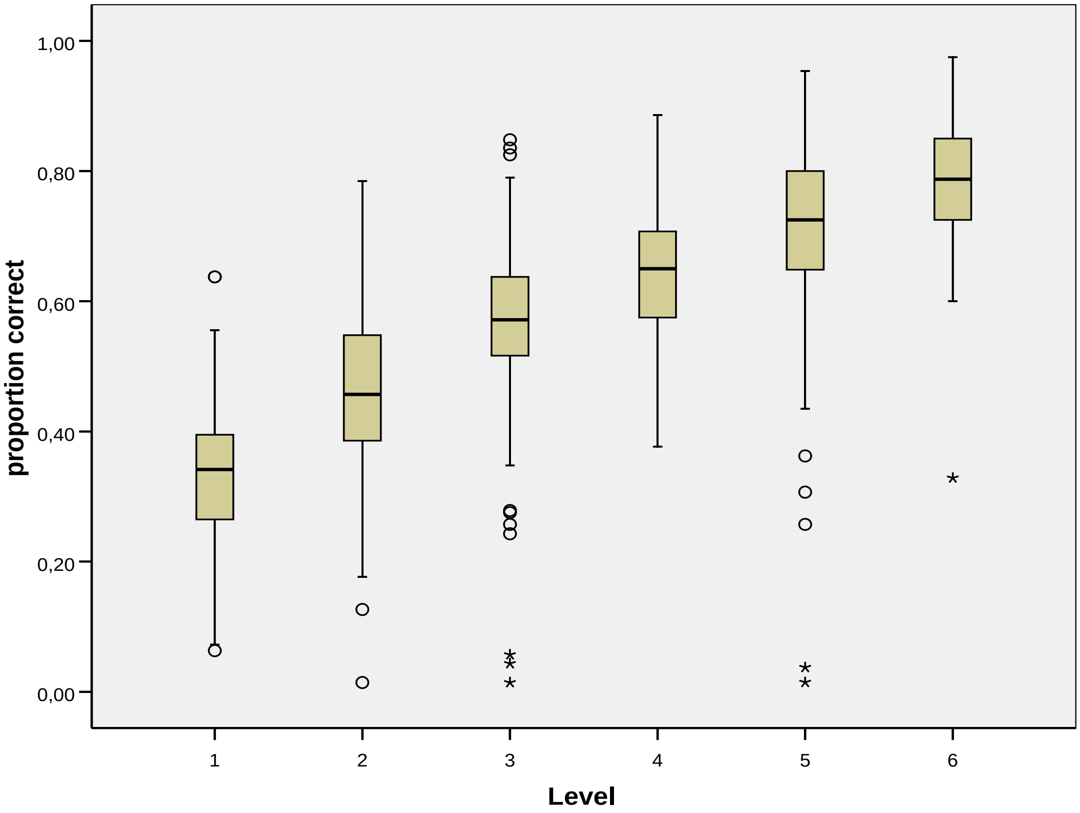
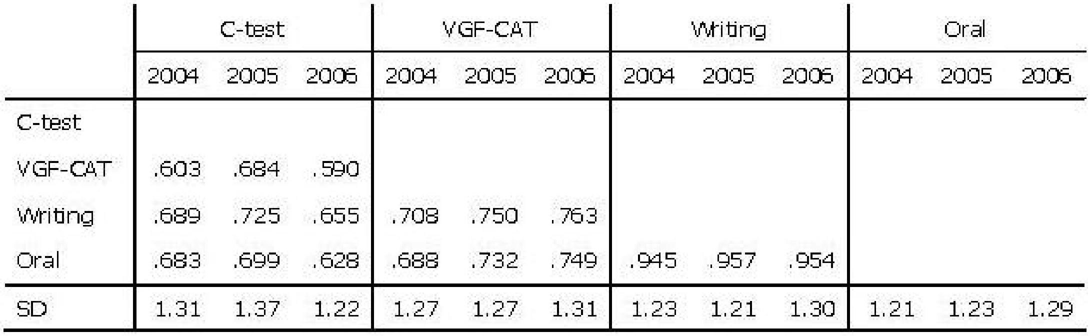


Figure 7. Box-Plot of Proportion of Correct Responses According to Course Level of Students



Pearson correlations between different certification components for results in 2004, 2005 and 2006 were calculated for the 6 course levels (see Table 2). The number of examinees varied between 802 and 806 for VGF scores, and 645 for the 2006 C-test.

Table 2. Pearson Correlation Coefficients Between Certification Components of SIMTEST and Their Standard Deviations (SD)



	C-test			VGF-CAT			Writing			Oral		
	2004	2005	2006	2004	2005	2006	2004	2005	2006	2004	2005	2006
C-test												
VGF-CAT	.603	.684	.590									
Writing	.689	.725	.655	.708	.750	.763						
Oral	.683	.699	.628	.688	.732	.749	.945	.957	.954			
SD	1.31	1.37	1.22	1.27	1.27	1.31	1.23	1.21	1.30	1.21	1.23	1.29

All correlation coefficients were statistically significant. As expected, they were very high between oral and writing components (r > 0.94) and moderate elsewhere (0.59 to 0.76) with the majority of values close to 0.70. "It is not very common for there to be no correlation between the results of two language tests. Since the different components are intended to test aspects of the same trait—language ability—they might be expected to show at least some degree of agreement".(Alderson et al, 1995, p. 78) According to the same authors, a correlation of + 0.70 would indicate "quite a strong agreement between scores". (p.79)

While correlations for VGF-CAT indicate stability over the three years analysed, those for the C-test diminished in 2006. This may be related to restriction of range because, as of 2006, the C-test is no longer a required component for levels 1 and 2.

Pros and Cons of the Test

The information below represents a summary of views expressed in various evaluation-specific surveys with teachers and students over the past three years.

Some students have expressed a dislike for being assessed by computer and feel that the evaluation is inevitably less reliable or valid than that made by a teacher. Meanwhile, some teachers have expressed concern about the lack of control over test content.

In terms of positive feedback, the use of SIMTEST in placement testing is seen by the majority of students and teachers as appropriate, efficient, and objective. In placement and certification situations, teachers greatly appreciate the automatic scoring and immediate results, while most perceive high correlations between certification results and their own class-based evaluations.

There is the face validity issue of giving a proficiency test at the end of a course, when an achievement test might more reasonably be expected. This has led to the sensation that course content is not sufficiently reflected in test items and that students are being tested on things they either do not know or which are above their level. Some students see the test as being the same for all students, while in fact, being adaptive, it is unique for each individual. But, traditionally, students expect specific exams for the course level they have studied.

Other students have complained that they feel that not enough questions are being asked for the evaluation to be fair—an opinion that is shared by some teachers. Most teachers agree that a built-in minimum number of CAT items might help to improve the face validity of the test. Some teachers have also suggested that the algorithm should also consider content distribution when administering items, e.g. balance vocabulary items with grammar items.

Conclusions and Future Development

In addressing the comments above, the following conclusions and the modifications proposed here are based on the experience and data gathered over the last six years.

An automatic pre-testing element could be incorporated into the system so that a built-in minimum number of items must be reached, by exposing a number of uncalibrated items to examinees. In this way, face validity can be improved in terms of the perception that there is a higher minimum number of items. Once the new items have seen sufficient exposure, they can be analyzed for potential inclusion as newly calibrated items without the costly necessity at present—especially in terms of logistics—of organizing special pre-testing sessions for the purposes of calibration

Technically, the uncalibrated items could be included in one of two places. The first, at the head of the CAT, as the initial four items before the CAT algorithm begins, i.e., before the program routinely estimates ability and the associated error of estimate. This might diminish the risk of overexposure for items presented in this initial phase. If so, these items would have to be given a value based on expert judgment.

Another suitable place would be at the end of the CAT, i.e,. as additional items presented after the algorithm has reached its conclusion and in response to the number of items required to reach an agreed minimum. This fluctuating necessity may, in itself, cause a variety of problems relating to the selection and an even exposure of uncalibrated items, however.

In response to the teachers' feedback the algorithm can be adjusted to allow for content distribution, e.g,. by distributing vocabulary and grammar in a balanced (non-random) manner that can be seen by teachers to coincide more closely with their expectations of students' ability. Alternatively, as Rudner and Guo (2007) have suggested, providing that sufficient items are available across all levels within the separate content areas under consideration, separate CATS might be the solution, i.e., (in our context) separate vocabulary and grammar CATs.

The algorithm used in the present measurement model is brusque in its oscillation between complete levels of ability. The test has, until its version 2.0, been calibrated according to CTT, not out of choice or design, but due to a lack of sufficient data for IRT calibration. After six years of an operational SIMTEST we now hope to fine-tune the algorithm and the test, via re-calibration of all items based on IRT and the incorporation of the points mentioned above, as SIMTEST 3.0.

In the light of continuing false impressions about the functioning of the CAT and concerns expressed there is an evident need and a sincere commitment to improve the quality of communication between all actors: testers, teachers and examinees.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University

Alderson, J. C., Clapham, C., Wall, D. (1995). *Language test construction and evaluation*. Cambridge, UK: Cambridge University

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge, UK: Cambridge University Press.

Dunkel, P. (1999). Research and development of a computer-adaptive test of listening comprehension in the less-commonly taught language Hausa. In M. Chalhoub-Deville (ed). *Issues in computer-adaptive testing of reading proficiency*. Cambridge, UK : Cambridge University Press.

Henning, G. T. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, CA.: Newbury House Publishers.

Rudner, L.M. & Guo, F. (2007) *A practitioner's perspective on computerized testing*. In Weiss, D. J. (2007). Proceedings of the 2007 GMAC conference on computerized adaptive testing. Available at www.psych.umn.edu/psylabs/CATCentral/

Zimowski, M.F., Muraki, E., Mislevy, R.J. & Bock, R.D. (1996). *BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International.