GFL
German as a foreign language

German Computer Adaptive Language Tests - Better Late Than Never

Felicitas Starr-Egger, Manchester

ISSN 1470 – 9570

German Computer Adaptive Language Tests - Better Late Than Never

Felicitas Starr-Egger, Manchester

Advances in IT have had a serious impact on language testing and led to the development of a large number of testing projects which utilise computers to varying degrees. The article briefly reviews computer adaptive language tests (CALTs) developed over the last 30 years, followed by an outline of the current situation with regard to German CALT projects. Subsequently potential key reasons for the apparent shortage of German test of this type are suggested. The issue of non-standardised terminology is discussed with reference to commonly used acronyms and conflicting definitions within literature. The article locates the area of computer adaptive testing (CAT) within computerised assessment of languages and concludes with a short introduction to the key features of a CAT and the underlying mathematical theory.

Key words: assessment, computer adaptive testing, computer adaptive language testing, DaF.

1. Projects in Computerised Language Testing

Assessment is probably the only area hated by learners and teachers alike. Pupils and students dread having to sit tests and exams and teachers dread having to set and mark them. The use of IT can be of benefit to all parties, especially in terms of making assessment more efficient, reliable and objective (Bull 1994, King 1995 and Callear and King 1997) as well as producing results more quickly (Mulkern 1998) and individualising tests (Henning 1991).

Within the field of computerised language testing, the concept of computer adaptive language tests (CALTs)[1] has recently received renewed attention. Probably the first fully adaptive language test was designed and used at Brigham Young University (Madsen 1991). Table 1 provides a few examples of the languages and skills covered (Schonenberg et al. 1993, Troubleyn et al. 1996, van Walle et al. 1996, Burston and Monville-Burston 1995, Madsen 1991, Stevenson and Gross 1991). Additional information was also retrieved through the ERIC and ERICAE online databases as well as several websites relating to projects described below.

The rapid growth of foreign language tests alongside projects initiated for English is worth noting and, more recently, researchers have started on projects for less commonly taught languages (see

[1] See section 2 for a terminology review.

Dunkel 1999b). It could be argued that due to the significant progress in this field over the past few years, CALT has established itself as an internationally recognised form of assessment.

Table 1 gives a few examples of adaptive tests developed from the early seventies onwards. However, obtaining precise facts and figures for some of the projects was hampered by a lack of published information. Noticeable is that most CALTs consist of multiple choice items, with item banks ranging from 35 to 40,000 entries, covering the four traditional skills (although not all projects include all skills) as well as grammar and/or vocabulary, and that furthermore most of the projects were developed outside Europe.

In the grid below the following abbreviations are used for item types: mc - multiple choice, match - matching answers, mr - multiple response

Table 1: Adaptive projects

Year	**Organisation**	**Title**	**Language**	**No. of items/ type**	**Skills**	**Comments**
1972-	ETS	TOEFL CBT	English	mc/match/mr	All	Linear in parts
1985	BYU	Yescat (Computest ESL)	English	608 (750)	Reading/ Listening	
1986	UCLA	ESL	English	–	All	
1988	Bur. Educ. Eval/NY univ.	(ACTFL/ILR)	French	100 texts 600 items	Reading	
1988	BYU	G-Cape/S-Cape (also Russ/Fr/Sp)	German/ Spanish	– mc	Reading, grammar	
1991	Didascalia	Atlas	French/ Dutch	40,000 –	Various	
1991	Montgomery C. Publ.Sch.	MEC	English	170	Reading/ Writing	
1994	Didascalia	Ariane	French/ Dutch	329 texts 3375 mc	Various	Final version not adaptive
1995/ 96	U/Melbourne	FrenchCat	French	125 mc	Reading	
1995	U/Melbourne	–	Japanese	225 mc	Grammar	
1996	CARLA- U/Minnesota	CoRa/CoSa/ CoWa/CoLa	Fr/Ger/Sp	35 mc	All	
1999	Ohio SU	–	Chinese	mc	Reading	
1999	Ohio SU	MultiCat	Fr/Ger/Sp	mc	Read/List/ Grammar	
1999-	EU - Socrates	Dialuc/Dialang	14 langs.*	Varied- mc/ short answers	All	Web-based
2000-	UCLES	BULATS	Eng/Fr/Sp/ Ger	Various	List/Read/ Grammar/ Vocab	
2000-	TestDaf Konsortium	TestDaF	German	various	All	Web based adaptive in parts

** Dutch, English, Finnish, Italian and Spanish to be available from February 2001*

A quick glance at the above table shows that even amongst the short selection of adaptive tests there is firstly an obvious lack of German and secondly projects including German only started in the late 80s. What complicates any attempt of an in-depth study of the field is the limited amount of information available on the German projects or the German part of the project. This may partly be due to their purpose, as it is not possible to gain access to confidential tests.

The following paragraphs outline the structure of some of the German tests.

- The CARLA website (which includes a link to a computer demo) provides several pages with test specifications on the Contextualized Speaking Assessment (CoSA), Contextualized Writing Assessment (CoWA), Contextualized Reading Assessment (CoRA) and the Contextualized Listening Assessment (CoLA), outlining the reasoning behind the project and the background to the Minnesota Language Proficiency Assessment (MLPA), and stating that the test covers reading, writing, speaking and listening. There is some content information, e.g. that CoRA consists of 35 multiple choice items and contains on average 12 texts with an average of three questions per text, but the level of detail varies between the individual skills pages. There is no reference to the underlying mathematical model.
- The Dialang website offers extensive information on the project management and partners. However there is no mention of the item bank size and unfortunately no details of its content. The skills reading, writing, listening, grammatical structures, and vocabulary will be covered. The project is currently in its piloting phase, a fully operational version is to be launched by December 2001 to coincide with the end of the European Year of Languages. Provided sufficient numbers of candidates take part in piloting the items all the 14 DIALANG languages should then come on-line. The items will be mainly multiple choice and short answers, and the underlying model is the Rasch[2] one-parameter model.

[2] This particular type of mathematical model was developed by the Danish mathematician Georg Rasch (1960) and expresses the probability of a correct response to an item as a function of a candidate's ability (θ) and the item's difficulty (b). A good introduction to some of the concepts behind the Rasch model can be found in Wright (1977) and Hambleton et al. (1991).

- The BULATS site describes the tests in slightly more detail, the computerised part (link to computer demo at the bottom of the web page) is described as an adaptive test and covers the same skills as the standard test.

  The standard 90 minute test covers listening (understanding short extracts and taking down phone messages), reading (understanding notices and short extracts, finding information and understanding a longer text) and grammar and vocabulary (gap filling and correction of errors by candidate). There is no information on the underlying mathematical model or the size of the item bank.

- The web site on TestDaf describes this test as comparable to IELTS and TOEFL, consisting of four skills and ultimately an adaptive test based on Rasch analysis is envisaged. The test covers the following skills:

  Reading (60 minutes): candidates are presented with short texts and are required to extract main points as well as specific information and to understand opinions.

  Listening (30 minutes): candidates hear a dialogue and are required to carry out the same tasks as for reading.

  Writing (60 minutes): candidates are asked to write two types of text. On the basis of given diagrams, graphs and tables they are asked to describe and compare. The second part is based on short texts, and candidates are required to contrast varying points of view on a topic as well as outline their own stance.

  Speaking (30 minutes): candidates demonstrate oral proficiency in four areas: making a request, situational speech acts such as giving/asking for information, describing and arguing.

  Unfortunately the website does not give any information on the size of the item bank. However, the site does cite a few publications in German, some of which are accessible through the project website and this might increase interest in computer (adaptive) tests in general within the DaF community.

So although computerised adaptive testing has now been used for decades in some areas, such as intelligence testing, its application to language testing, whilst regarded as a valid method, still lacks the penetration computer-assisted testing in general has achieved[3]. It is possible to identify 4 key reasons why there are so few German CAT projects, the first being the rather complicated mathematical background to adaptive testing. Baker (1997) points out that there has been an unusally long gap between the development of the mathematical models used and their application to language testing, and that this in turn was and still is due to the inaccessibility of the mathematics and statistics underpinning the notion of CAT, but also because of what Baker (1997: 23) calls the "paucity of articles explaining them (= the mathematical models) to practitioners".[4]

The second reason is the problem of the large number of candidates required during CAT development[5] which may prove a considerable obstacle, especially for languages such as German, given the steady decrease in student numbers. Here again, the possibility of trialling the tests via the web will make a considerable difference.

CALTs in particular, as well as CATs in general, only seem to flourish in an environment which uses standardised tests for decision making processes, such as admission to university courses or army recruitment or immigration. Intelligence testing used for the selection of army recruits is one example (Grist et al. 1989, Sands et al. 1997), but there are many others such as the GRE (the Graduate Record Examinations used in the USA for college admission to postgraduate courses) or ATLAS (the project designed to test Belgian civil servants' second language proficiency) (see table 1). Here we find the third reason why up to now there had been an almost complete absence of German CALT projects in Europe, as there had been no "official", either government driven or university driven, interest in the development of such a testing mechanism. It was only when German universities noticed a drop in applications from foreign students and the subsequent discussion about the "Studienstandort Deutschland" prompted an inquiry into the possible causes that the need for a truly standardised German test became apparent.[6]

[3] See also Fairtest (1998).

[4] Henning (1987) and McNamara (1996:149ff and 2000) provide excellent introductions to the subject for language teachers.

[5] For a discussion of the problem of calibration sample size, see Lord (1983), Henning (1988) as well as Linacre (1994).

[6] See Gutzat et al. (2001) for the background to TestDaF.

Finally, generating computerised tests is time and man-power intensive,[7] therefore the question of project finance is not to be underestimated. For TestDaF, work commenced in 1998 and the first official test use was scheduled for 2001; development of the six modules of ATLAS “took 14 man-years, and was realised in five years”[8]. This also means that projects need the financial backing of large organisations, for example Dialang is funded by the EC and TestDaF by the German Foreign Office and the Ministry of Education and Research. The fourth reason for the lack of German CATs is therefore a lack of funding.

Dialang and TestDaF are web-based and therefore accessible to a large number of practitioners and researchers alike, moreover, the former covers 14 languages, thereby hopefully generating interest in this type of testing within the international foreign language teaching/researching community.

Apart from the key development issues mentioned above, one could also list a minor factor as potentially hampering research and possibly progress within computerised language testing: inconsistent terminology. The following section provides an outline of acronyms used.

2. Types and terminology

The terminology of computerised assessment followed the increase in the use of computers in teaching and learning in general: the next logical step was to complete this process and extend their use to assessment as well (Madsen 1991: 240). However, in general the extent to which IT is used is not always clearly defined (Brown et al.1997: 204). Computerised assessment might mean that students submit essays to a publicly accessible directory and then the marked piece is returned by e-mail or it is left in the original directory for peer assessment. It might also mean that a multiple choice test is automatically checked and marked through the use of OCR software, or that a test is presented on computer, then printed, scored by hand and returned to students as a hard copy.

The next variation might be that the test is presented and scored by computer, but all candidates see the same questions, and finally all phases, including the choice of questions, could be fully

[7] See Dunkel (1997).

[8] Troubleyn et al. (1996:359).

automated, and also “tailored” (Dunkel 1997). This last variant is commonly described as computerised adaptive testing or CAT. Here, with the exception of the piloting phase, all stages in the preparation and testing procedure are exclusively carried out on and by computers.

In order to map the field of computerised assessment and arrive at a more precise location of CATs within it, it is necessary to take a closer look at some of the terms and acronyms used. However, since new terms are being added all the time, it is impossible to provide a truly comprehensive review. Furthermore, there is considerable confusion with regard to terminology in this area (Levy 1997: 80f), partly because some terms are more commonly used in the USA or Europe (Wyatt 1984 cited in Levy 1997) and partly because some emphasise one particular aspect (Levy 1997: 79).

The terms commonly used in this field are CBE (Computer Based Education), CBT (Computer Based Testing), CBELT (Computer Based English Language Testing), CAA (Computer Aided/Assisted Assessment), CAT (meaning either Computer Aided/Assisted or Computer Adaptive Testing), CALT (Computer Adaptive Language Testing). It seems appropriate firstly to isolate the area of Computer Based Education or CBE (Levy 1997: 77ff) within education in general. This immediately leads to the question why the term “based” is used, rather than “assisted”. Levy explains that “When the word ‘based’ is part of the acronym we are looking at a very central, all-encompassing role for the computer” (ibid., 78), whereas “the words ‘aided’ or ‘assisted’ … highlight the computer’s subservient, auxiliary role” (Ahmad *et al*. 1985: 2). This could mean that CBE includes types of computerised instruction as well as computerised assesment, where every step of the process (item selection, scoring, feedback) is carried out by computers, whereas in CAA computers only carry out certain parts of the process. One might argue that whilst ‘based’ is an appropriate term as such, it might be preferable to ascribe it a more general meaning, i.e. denoting all forms of teaching and assessment which involve a computer in the widest sense, subsuming all other specialised forms.

CAA is generally understood as computer assisted assessment, defined as the use of IT for the purpose of assessment (Bull 1994, King 1995 and Callear and King 1997). What is important is the fact that for the UK Centre for Computer Assisted Assessment areas such as the collation and analysis of data gathered belong to this area as much as the delivery and marking of exams (http://www.caacentre.ac.uk/).

The Centre gives a contrasting definition of the term “based” in the context of CBA meaning computer based assessment “in which the questions or tasks are delivered to a student via a computer terminal”, a very general definition which does not indicate how the scoring is carried out (www.caacentre.ac.uk/fqgen2.shtml) but one might, in other words, say that here CBA is a form of CAA. This does not quite match Levy’s interpretation of the term “based”. To confuse the issue even further, the UCLES website offers adaptive tests under the CBA banner (http://www.computer-based-testing.org/).

Instead of the above-mentioned broad definitions, it would be helpful if terms could be associated with the extent to which computers are employed in the process and the terminology standardised. I would therefore like to suggest the following:

a) “based” refers to the fact that either item presentation and feedback or scoring are computerised,

b) “aided” or “assisted” mean that item presentation, feedback and scoring are carried out by the computer, and that

c) “adaptive” refers to the only type where specifically item selection, presentation, scoring and feedback are computerised.

When trying to understand the distinctions between CAT and CALT, one encounters an even greater terminological tangle. In the literature the acronym CAT is not only used for computer assisted test or testing (Madsen 1991) but also for computer adaptive testing (Tung 1986, Canale 1986, Dunkel 1991, Noijons 1994). Given that an “adaptive” procedure has unique characteristics (described in detail below) not shared with “ordinary” computer assisted test routines, a clear differentiation between the two is imperative.

A look at the term CALT shows that, again, different authors interpret this acronym in different ways. For Noijons this refers to “any type of computerized language test” (1994: 43), where CAT denotes the adaptive variant. Meunier (1994), however, interprets CALT as “adaptive”. Whilst the seasoned researcher and language tester may be able to deal with this quite comfortably, any novice in this field would probably find the lack of uniformity quite confusing.

Two terms still remain, CBELT and CBT. Madsen (1991) tries to distinguish between CAT (interpreted as assisted and denoting any form of computerised test) and CBELT. With reference to Alderson (1987), the latter is described as a CAT variant where scoring is carried out by the computer but the test is not adaptive. Dunkel (1999a) uses CBT meaning a fixed, linear test, a type that was above covered by the first definition of the term CBA.

Table 2 below summarises the terms used in the literature. They are shown in relation to the categories "adaptive" and "linear" (meaning that the items shown to the candidates are pre-selected).[9]

Table 2: Summary of terms

Acronym	Meaning	"Adaptive"	"Linear"
CBE	Computer Based Education		✓
CBT	Computer Based Testing		✓
CBA	Computer Based Assessment	✓	✓
CBELT	Computer Based English Language Teaching		✓
CAA	Computer Aided/Assisted Assessment		✓
CAT	Computer Assisted Testing Computer Adaptive Testing	✓ ✓	✓ ✓
CALT	Computer Assisted Language Testing Computer Adaptive Language Testing	✓ ✓	✓ ✓

One can see that they are all used by certain authors to refer to a linear assessment method and some are also employed to denote the special case of computerised testing which is self-adjusting to the student's ability.

There is no simple solution to this problem, as it is not a case of say European test developers and researchers using one set of terms and American using another, yet clear equivalents can be established. No doubt this confusion will continue into the future, certainly until terminological standardisation is agreed, which in this case would constitute a great step forward.

For the remainder of this article the terms will be understood as follows. CAA denotes any type of assessment using a computer. CAT and CALT refer to an adaptive procedure, however CAT

[9] The term is slightly misleading. The key idea here is that all candidates must work through the same number of items, irrespective of the order of presentation.

could be used for any field, whereas a CALT is specifically a language test. This corresponds to CAL and CALL. The following section now describes the "adaptive" procedure in more detail.

3 Theoretical Aspects of CATs

Rapid progress in computing over the last two decades, rekindled testers' and language testers' interest in all types of computerised testing and revolutionised its practical aspects. With regard to CATs the calculation methods underpinning the procedure as well as the administration of the tests as such experienced unprecedented progress.

Noijons (1994: 38) defines adaptive testing as an integrated procedure in which language performance is elicited and assessed with the help of a computer, consisting of three integrated procedures:

1. generating the test
2. interaction with candidate
3. evaluation of response.

Apart from these, any test needs an item bank, a pool of "questions". The main distinguishing feature of a CALT, however, is the fact that the test is "adaptive", in other words the test adjusts itself to the candidate's ability and depending on whether a correct or incorrect response is given, the next item is more difficult or easier than the last one. In traditional tests, all candidates must attempt all questions which are presented in a predetermined sequence. In "adaptive" tests an algorithm chooses items from an item bank which match the candidate's ability level. This is possible because as the result of a pilot study, the items have been ascribed a certain difficulty level.[10] CALTs typically use multiple choice items, cloze or scrambled sentences.

CATs and CALTs are founded on the theory that a candidate's observed behaviour (response) when encountering an item is based on a latent ability (trait). These "traits" are not directly observable attitudes and abilities. According to the underlying theoretical framework, Item

[10] See Wright (1968) and Wainer et al. (1990) for a description of the mathematical procedure.

Response Theory[11], only one trait should be measured (Madsen 1991: 239). Meunier (1994) points out that it may be difficult to construct a test to assess reading or communicative skills, since these consist of several aspects.

Computer adaptive testing can also be described as “tailored testing” (Dunkel 1997, Madsen 1991: 238). Unlike in a conventional test where all candidates see the same number of items, albeit not necessarily in the same sequence, here the testee is presented with a selection of items from an “item pool”, a bank of questions. The “test takers respond to test items, a computer-adaptive test “adapts” itself to test takers by selecting the next item to be presented on the basis of performance on preceding items” (Mulkern 1998).

A correct answer leads to an increased level of difficulty, while an incorrect answer lowers the level of difficulty. The computer administers the test according to pre-set increments or decrements, in other words the test designer can decide how big the steps between presented items are. In S-Cape, a Spanish test, for example, the increments were set at 6 for a correct response and the decrements set at 5 for an incorrect one (Larson 1987, 1988 cited in Madsen 1991). A different approach would be to estimate a candidate’s proficiency after each response and choose an item that would yield the most information (Wainer et al. 1990: 128).

The concept of adaptive testing can be likened to the procedure in an oral exam, in which the examiner responds to the candidate’s answers (ibid: 10) by adjusting the level of subsequent questions according to the previous answer. If the ultimate aim of the exam is to obtain a precise picture of the testee’s ability, no information can be gained about the candidate if the questions asked are too hard or too easy.

The question of when the test should stop has turned out to be rather difficult. It might end once a predetermined number of items has been administered or after a certain time has elapsed. Experts appear to agree that the test should stop once an acceptable level of accuracy has been reached (Henning 1988: 132, Kaya-Carton et al. 1991), however the term “acceptable accuracy” itself is open to debate. It could be defined by the standard error of measurement and this should ideally

[11] For a detailed discussion of how to construct a CAT and a very readable introduction, see Wainer et al. (1990). Hambleton and Swaminathan (1985) is widely quoted as the best introduction to the mathematics behind IRT, although the not so mathematically inclined may find Hambleton et al. (1991) more digestible.

be 0. The problem arises from the fact that "the standard error of measurement decreases in inverse proportion to the number of questions presented" (Burston and Monville-Burston 1995: 35) and since the test should be as short as possible a compromise has to be sought. Essenius (1995: 22) describes the use of the item pool score (θ_c), i.e. "the expected score of a student with a certain proficiency, if given all items in the pool" to decide the cut-off point for a mastery test. Once the confidence interval for the student's proficiency level no longer includes θ_c the test stops. If the confidence interval is above θ_c, mastery is concluded, if it is below non-mastery is assumed. Thissen and Mislevy (1990) argue in favour of a stopping rule based on a certain precision level of the student's ability θ. In the example given, the stopping criterion, the standard error, was set to a maximum and once this value is reached the test is terminated.

In general all tests must meet general criteria (reliability, validity, objectivity and ability to discriminate). There are numerous publications on this issue. Henning (1987) covers these but also provides an excellent introduction not only to language testing in general but to IRT and adaptive testing. A considerable amount of research effort is directed towards developing the theoretical base as well as the practical approaches needed to verify that tests actually meet the above-mentioned requirements.

As mentioned above, the test draws items from an item bank. During the development of any CAT project by far the most effort is directed towards the compilation and calibration of this item bank. All CAT projects therefore consist of two phases: the initial item calibration phase, during which a first value for item difficulty is obtained, and the second, actual testing phase, which not only provides information on the students' ability but also verifies the item parameters.

The principles of CAT have been applied to various subject areas, mainly to intelligence testing (Sands et al. 1997) but also to engineering EEE/TAIGA (Essenius 1995) or mathematics as in TIMMS (1996) and, of course, extensively to English language testing.

CATs could be summarised in the following list of advantages and disadvantages.

Advantages

There are some advantages CAA and CAT have in common (Grist et al. 1989 Madsen 1986, 1991, Callear and King 1997). Among them are the option to "receive immediate feedback on the students' performance" (Mulkern 1998). This in turn not only enhances motivation levels but also improves the learning experience provided by the test designer.

Universities and colleges are usually equipped with large computer clusters. For tests under such conditions, scheduling and supervision concerns are greatly reduced because individual administration is possible (Mulkern 1998).

Data storage can also be improved, some authoring packages provide links to databases for faster and more efficient processing of results.

There are also cost savings to be gained through CAA and CALT. These may be calculated in different ways (Callear and King 1997):

a) In the conventional methods different tests had to be prepared for different groups of students. This is very time consuming for the lecturers involved and also requires resources such as paper, printing and photocopying.

b) Marking (and second marking) are even more time consuming than the preparation stages. Here the use of computers not only reduces the amount of time spent on results, it also guarantees consistency.

As far as CATs are concerned, candidates develop a more positive attitude towards the test. The fact that the items presented are well matched to the candidates' ability improves their attitude towards the test. The test taker is continuously faced with a realistic challenge - items are not too difficult or too easy (Mulkern 1998). CAT technologies have also been found to improve test-taking motivation and to reduce average test score differences across ethnic groups.

A large number of items provides matches to a larger range of personal abilities (Henning 1991, Mulkern 1998) while at the same time yielding more accurate results.

Security is an important issue with regards to computerised testing. However, it is not only the test answer data that needs to be protected from unauthorised access, the test files as such may be equally at risk. Students may be tempted to try and obtain the questions and potentially prepare an answer file in advance. For computer-assisted assessment such a scenario is feasible[12], but for adaptive testing any attempts to cheat are usually futile. Even a moderately large item bank

[12] See King et al. (1998) on security issues involved and attempts to deal with them.

combined with a selection algorithm make it virtually impossible for students to cheat (Madsen 1991: 241), since the item sequence cannot be predicted. Therefore, because each test taker is administered a different set of test items, test security is enhanced (Mulkern 1998).

It should be stressed that this improved security[13] can only be guaranteed if test items are drawn from a large item bank and that no two students are presented with the same tests.

CATs are usually shorter than conventional pen-and-paper versions, since only a relatively small number of items has to be presented for the candidate's ability to be determined (Mulkern 1998, Stephens 1994).

Disadvantages

The following points are commonly cited as the main disadvantages: lack of uniform time distribution across items, unidimensionality, the fact that the development of tests is time consuming, student unfamiliarity with medium, validity, absence of communication (Madsen 1991, Henning 1991, Canale 1986, Tung 1986).

The first point is easily misunderstood since the fact that students spend varying lengths of time on items is not only true for CATs but for all tests. There is a fundamental difference between conventional tests and CATs. Traditional tests finish once a certain time has elapsed or the candidate has attempted all questions, while CATs finish once a pre-determined stopping criterion has been fulfilled. Furthermore, different students reach this point after varying time spans. So the criticism centres on the question of how it is possible to obtain an accurate measurement under different testing conditions. CATs achieve this, because the stop criterion is the same for every test candidate and if the stop condition is for instance a precision indicator such as the standard error, the time factor is not relevant. Of course other stop criteria could also be applied (cf. Rudner 1998).

As far as unidimensionality is concerned, this again could be seen as a positive feature, since results from different tests could be combined into a more accurate student profile, enabling teachers and students to focus better on areas in need of additional work.

The criticism that CAA and CAT development is time consuming has to be accepted, yet it should be borne in mind that long term savings might be gained (Callear and King 1997). To be precise, different phases in the development of a CAT require varying amounts of time, depending on how many people or groups are involved. In general it is probably the calibration phase which takes longest.

Students' unfamiliarity with IT had considerable impact until PCs became a more commonplace teaching and learning tool and students developed a more positive attitude towards computers (Reid 1986; Neu and Scarcella 1991; Phinney 1991 all cited in Brow n 1997).

The issue of validity (as well as reliability) has been investigated quite extensively (Steinberg et al. 1990; McBride and Martin 1983; Bennet and Rock 1995 cited in Brown 1997 and Thissen and Mislevy 1990) and continues to be the subject of research. Dunkel (1999a) points out that validity issues exist not only with regard to CATs but also for item response theory on which they are based.

The criticism that CATs are non-communicative, in the sense that the presentation of a text followed by multiple choice questions is not an authentic interaction is valid, however the potential use of multimedia and speech recognition in testing gives computerised testing in general an advantage over paper-based methods. Brown (1997) and Dunkel (1999a) admit that research in the area of multimedia and speech recognition is intensive in terms of expense, labour and processing power required.

Conclusion

Despite the fact that the concept of adaptive testing has been around for several decades, it did not generate much interest within the language teaching and research communities for a long time. In the case of German, there are still very few projects and literature is generally difficult to obtain. This is due to several factors: the absence of the need for a truly standardised test, the high cost, the long development time for CALTS, and the large number of candidates required during the piloting phase. One additional shortcoming of the field of computerised language

[13] Brown (1997) raises a few general logistical questions for CALT and some security issues in relation to item banks arising from American legislation on the disclosure of tests.

testing, is the considerable inconsistency with regard to terminology. This may constitute an additional obstacle for the novice language tester in particular in adaptive testing, which due to its psychometric and therefore mathematical basis, already presents a challenge to language practitioners. Initiatives such as the *Dialang* project or *TestDaF* will hopefully generate more interest and enthusiasm for this type of assessment within the *DaF* community.

Bibliography

Ahmad, K.; Corbett, G.; Rodgers, M.; Sussex, R. (1985) *Computers, Language Learning and Language Teaching*. Cambridge: Cambridge University Press

Alderson, Charles (1987) Innovations in Language Testing: Can the microcomputer help? Special Report No. 1: *Language Testing Update.* Lancaster, UK: University of Lancaster

Baker, Rosemary (1997) Classical Test Theory and Item Response Theory in Test Analysis. *Language Testing Update*. Special Report No.2

Bennett, R. E.; Rock, D. A. (1995). Generalizability, validity, and examinee perceptions of a computer-delivered formulating-hypotheses test. *Journal of Educational Measurement* 32, 19-36.

Brown, James D. (1997) Computers in Language Testing Present Research and Some Future Directions. *Language Learning & Technology* 1(1), 44-59

Brown, G.A.; Bull, Joanna; Pendlebury, M. (1997) Assessing Student Learning in Higher Education. London: Routledge

Bull, Joanna (1994) Computer based assessment some issues for consideration. *Active Learning* 1, 18 - 21

Burston, Jack and Monville-Burston, Monique (1995) Practical Design and Implementation Considerations of a Computer Adaptive Foreign Language Test The Monash /Melbourne French CAT. *CALICO Journal* 13(1), 26-46

Callear, David and King, Terry (1997) Using Computer Based Tests for Information Science. *ALT-J* 5(1), 27-32

Canale, Michael (1986) The Promise and Threat of Computerized Adaptive Assessment of Reading Comprehension. In: Stansfield, Charles W. (ed.) 29-46

Dunkel, Patricia (ed.) (1991) *Computer-Assisted Language Learning and Testing. Research Issues and Practice*. New York: Newbury House.

Dunkel, Patricia (1997) Computer-Adaptive Testing of Listening Comprehension. A Blueprint for CAT Development. *The Language Teacher Online*. Retrieved from http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97/oct/dunkel.html

Dunkel, Patricia (1999a) Considerations in Developing or Using Second/Foreign Language Proficiency Computer-Adaptive Tests. *Language Learning & Technology*; 2(2), 77-93

Dunkel, Patricia (1999b) Research and development of computer adaptive test of listening comprehension in the less-commonly taught language Hausa. In: Chalhoub-Deville, M. (ed.) 91-118.

Essenius, Rik (1995) *Adaptive Computer-Based Training in Engineering Education*. Proefschrift Technische Universiteit Delft

Fairtest (1998) *Computerized Testing: More Questions Than Answers.* FairTest Fact Sheet. http://www.fairtest.org.

Grist, Susan et al. (1989) *Computerized Adaptive Tests*. ERIC Digest No. 107. Retrieved from http://ericae.net/db/edo/ED315425.htm

Gutzat, Bärbel (2001) Computer- und Interneteinsatz bei TestDaF. *Veröffentlichungen zu TestDaF*. Retrieved from http://www.testdaf.de/veroeffentlichungen/veroeff.html

Hambleton, Ronald; Swaminathan, Hariharan (1985) *Item Response Theory. Principles and Applications*. Boston: Kluwer Nijhoff Publishing

Hambleton, Ronald; Swaminathan, Hariharan; Rogers, H. Jane (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage

Henning, Grant (1987) *A Guide to Language Testing*. Cambridge: Newbury House

Henning, Grant (1988) Computer Adaptive Language Testing: An Algorithm for Item Selection. In: Jung, Udo O.H. (ed.), 129-135

Henning, Grant (1991) Validating an item-bank in a computer-adaptive test using Item Response Theory in the process of validating CATS. In: Dunkel, P. (ed.), 209-222

Kaya-Carton, Esin; Carton, Aaron S.; Dandonoli, Patricia (1991) Developing a computer-adaptive test of French reading proficiency. In: Dunkel, P. (ed.), 259-284

King, Terry (1995) Using Computer Aided Assessment for Objective Testing in Higher Education. A Case Study at the University of Portsmouth, *SEDA Paper 88, Innovations in Computing Teaching*, 123-128

King, Terry; Billinge, D.; Callear, D.; Wilson, S.; Wilson, A.; Briggs, J. (1998) *Developing and Evaluating a CAA Protocol for University Students*. Retrieved from http://www.dis.port.ac.uk/~kingtr/research/sacaa.html

Larson, J. W. (1987) Computerized Adaptive Language Testing: A Spanish Adaptive Placement Exam. In: Bailey, K.M. et al. *Language Testing Research*, 1-10

Larson, J. W. (1988) S-CAPE: A Spanish Computerized Adaptive Placement Exam. In: Flint Smith, William *Modern technology in foreign language education; Applications and projects.* ACTFL Foreign Language Education Series, 277-289

Levy, Michael (1997) *Computer-Assisted Language Learning. Context and Conceptualization*. Oxford: Clarendon Press

Linacre, John (1994) Sample Size and Item Calibration Stability. *Rasch Measurement Transactions* 7 (4), 328-329.

Lord, Frederic (1983) Small *n* Justifies the Rasch Model. In: Weiss, David J., (ed.) (1983). *New Horizons in Testing*. New York: Academic Press, 51-61.

Madsen, Harold (1986) Evaluating a computer-adaptive ESL placement test. *CALICO Journal* 4 (2), 41-50

Madsen, H. (1991) Computer-adaptive testing of listening and reading comprehension. The Brigham Young University approach. In: Dunkel, P. (ed.) (1991), 237-257

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In: D. J. Weiss (ed.), *New horizons in testing: Latent trait test theory and computerized adaptive testing*, New York: Academic Press, 223-236.

McNamara, Tim (1996) *Measuring Second Language Performance*. London: Longman.

McNamara, Tim (2000) *Language Testing*. Oxford: Oxford University Press.

Meunier, Lydie E. (1994) Computer Adaptive Language Tests (CALT) Offer a Great Potential for Functional Testing. Yet Why Don't They? *CALICO Journal* 11(4), 23-39.

Mulkern, A.E. (1998) *Frequently Asked Questions about Computer-Adaptive Testing (CAT)*. Retrieved from http://carla.acad.umn.edu/CATFAQ.html (CARLA/Institute of International Studies and Programs/Uinversity of Minnesota).

Neu, J., Scarcella, R. (1991) Word processing in the ESL writing classroom: A survey of student attitudes. In: Dunkel, P. (ed.), 169-187.

Noijons, José (1994) Testing Computer Assisted language Tests. Towards a Checklist for CALT. *CALICO Journal* 12(1), 37 – 58.

Phinney, M. (1991) Computer assisted writing and writing apprehension in ESL students. In: Dunkel, P. (ed.), 189-204.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. [Danish Institute of Educational Research 1960; University of Chicago Press 1980; MESA Press 1993] Chicago: MESA Press.

Reid, J. (1986) Using the writer's workbench in composition teaching and testing. In: Stansfield, Charles W. (ed.), 167-188.

Rudner, Lawrence M. (1998). *Item Banking*. Retrieved from http://ericae.net/digests/tm9805.htm

Sands, William A.; Waters, Brian K.; McBride, James R. (eds.) (1997) *Computerized Adaptive Testing. From Inquiry to Operation*. Washington, DC: American Psychological Association

Schonenberg, Nancy; Van Achter, Ilse.; Van Walle, A.; van Deun, K.; Decoo, Wilfred; Colpaert, Josef (1993) ATLAS An Adaptive Model for Measuring Language Proficiency. *CALICO Journal* 11 (2), 45-50

Stansfield, Charles W. (ed.) (1986) *Technology and language testing*. Washington, DC: Teachers of English to Speakers of Other Languages.

Steinberg, L.; Thissen, D.; Wainer, H. (1990). Validity. In H. Wainer, N. J. Dorans, R. Flaugher, B. F. Green, R. J. Mislevy, L. Steinberg, & D. Thissen (eds.), *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum, 187-232.

Stephens, Derek (1994) Using computer assisted assessment time saver or sophisticated distraction? *Active Learning* 1, 11-15.

Stevenson, J.; Gross, S. (1991) Use of a computerized testing model for ESOL, bilingual entry/exit decision making. In: Dunkel, P. (ed.), 223-235.

TIMMS (1996) *Third International Mathematics and Science Study*. Retrieved from http://www.mpib_berlin.mpg.de/TIMMS_II/Testaufgaben/Einfuehrung. html

Thissen, David; Mislevy, Robert J. (1990) Testing Algorithms. In: Wainer et al. (eds), 103-134.

Troubleyn, Katrien; Heireman, Kathleen; Van Walle, Anne (1996) ATLAS Computerised Second Language Proficiency Testing. *CALICO Journal* 9(4), 359-366

Tung, P. (1986) Computerized adaptive testing implications for language test developers. In: Stansfield, Charles W. (ed.), 12-28.

Van Walle, Ann; Heireman, Kathleen; Troubleyn, Katrien (1996) ARIANE. Computerised Mother Tongue Tests For Belgian Civil Servants, *CALICO Journal* 9(4), 367-374.

Wainer, Howard; Dorans, N.J.; Flaugher, R.; Green, B.F.; Mislevy; R.J. Steinberg, L.; Thissen, D. (eds.) (1990). *Computerized Adaptive Testing: A Primer*. Hillsdale, NJ: Lawrence Erlbaum.

Wright, Benjamin D. (1968). Sample-free test calibration and person measurement. *Proceedings of the 1967 Invitational Conference on Testing*, Princeton: Educational Testing Service, 85-101.

Wright, Benjamin D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, (14), 97-116.

Wyatt, D.H. (1984) *Computers and ESL.* Englewood Cliffs, NJ: Prentice Hall

Biographical information

Felicitas M. Starr-Egger is currently foreign language coordinator at the Language Teaching Centre at UMIST and working towards a PhD in computer-assisted testing at the Department of Language and Linguistics at UMIST. After completing the *Lehramt für Höhere Schulen* for English and German at the Universität für Bildungswissenschaften in Klagenfurt/Austria she worked as a technical translator and German teacher in Bristol and Manchester.