COMPUTER-ADAPTIVE TESTING IN HIGHER EDUCATION: THE VALIDITY AND RELIABILITY OF THE APPROACH

Trevor Barker

Computer Adaptive Testing in Higher Education: The Validity and Reliability of the Approach

Trevor Barker
Department of Computer Science
University of Hertfordshire, UK
t.1.barker@herts.ac.uk

Abstract

This paper presents a summary of a six year study into the design, implementation and evaluation of a computer-adaptive test (CAT) for the assessment of Computer Science undergraduates in a UK university. In the first part of this project, a series of empirical studies were carried out in order to evaluate the contribution that the CAT approach could make to the assessment of Computer Science undergraduates. A brief summary of this research is presented in this paper. It was found in this research that the developed CAT was effective at tailoring the level of difficulty of the test to the ability of individual students. The two main groups of stakeholders, students and academic staff, both exhibited a positive attitude towards the CAT approach and the user interface. In the main part of this paper, the validity and reliability of the CAT approach is assessed. Two empirical studies were undertaken in order test the CAT's validity and reliability and the results of these studies are presented here. Findings from this research are interpreted to show that in the context of assessment in Higher Education, the CAT developed in this research was valid and reliable. In the concluding section these findings are discussed in relation to other research in this area.

1.0 Introduction

In Higher Education today, increasing reliance is being placed upon the use of online systems for learning and assessment. At the University of Hertfordshire, Computer Based Testing (CBT) is used extensively for formative and summative testing on undergraduate modules. One of the main characteristics of a CBT is that all students are presented with the same set of predefined questions. This static approach, however, is likely to pose problems for individual students. For example, what might seem a difficult and therefore frustrating question to one student could seem too easy and thus uninteresting to another. A key factor in determining the usefulness of an assessment strategy is student engagement. If students are faced with assessment tasks that are de-motivating, the expected benefits of assessment could be lessened.

One potential way to address this issue would be the inclusion of computer-adaptive tests (CATs) as part of student assessment. In a CAT, the sequence and level of difficulty of the questions administered during the test are dynamically selected based on performance during the test. In other words, the proficiency level of individual students is estimated during the test so the questions can be tailored to match each student's proficiency level within the subject domain. Brusilovsky (2004) cites the CAT approach as one of the elements of a paradigm shift within educational software development, from "one size fits all" to one capable of offering higher levels of interaction and personalisation.

1.1 Computer-adaptive testing

Computer-adaptive tests (CATs) are based on Item Response Theory (IRT). IRT is a family of mathematical functions that attempts to predict the probability of a student correctly answering a question. The CAT software application developed at the University of Hertfordshire comprised a graphical user interface, a database of questions and an adaptive algorithm. The adaptive algorithm was based on the Three-Parameter Logistic (3-PL) model from IRT. Equation 1 shows the 3-PL function used to estimate the probability of a user with an unknown proficiency level θ correctly answering a question of difficulty *b*, discrimination *a* and pseudo-chance *c*.

Equation 1: The Three-Parameter Logistic Model

$$P(\theta) = c + \frac{1-c}{1+e^{-1.7a(\theta-b)}}$$

Equation 2: The Likelihood Function

$$L(u_1, u_2, ..., u_n \mid \theta) = \prod_{j=1}^{n} P_j^{u_j} Q_j^{1-u_j}$$

A typical CAT starts with a question of average difficulty. In general terms, a correct response will cause a more difficult question to be administered next. Conversely, an incorrect response will cause a less difficult question to follow. For each question answered, the mathematical function shown in Equation 1 is used to estimate the student's proficiency level. The question to be administered next as well as the final score obtained by any given user is computed using the Likelihood function shown in Equation 2. A fuller account of IRT can be found in Lord (1980) and Wainer (2000).

1.2 Development of the prototype

The development of the prototype employed in this research has been the subject of much research over a six year period. Empirical studies were undertaken in order to determine the most appropriate test conditions. Using an iterative prototyping software development method in conjunction, empirical studies were used to establish the following parameters for the CAT.

- Starting Condition (Lilley et al., 2004)
- Test duration (Lilley et al., 2004)

- Stopping condition: (Lilley et al., 2002; Lilley et al.,2004)
- Effect of question review (Lilley & Barker, 2005)
- Calibration of database: (Barker et al., 2006)

In addition, it was important to show that student and staff attitude to the CAT approach was positive. Several studies were undertaken to this end.

- Test-taker evaluation of the CAT approach (Lilley & Barker, 2003; Lilley & Barker, 2006)
- Academic staff evaluation of the CAT approach (Barker & Lilley, 2006)

In summary, this research was instrumental in setting up the test conditions for the CAT application and was able to show that the CAT approach was acceptable to students and staff both in a formative and summative context.

2.0 Validity and Reliability of the CAT approach

Validity and reliability are of crucial importance to all stakeholders in the student assessment process, including students, academic staff, educational institutions and prospective employers. In previous research, testing conditions for the CAT prototype were established and student and academic staff attitude towards, and acceptance of, the computer-adaptive test (CAT) approach were examined. Dunn and colleagues (2003: p17) caution us however, that “it is important for all stakeholders in the assessment process that the measurement of performance is valid and reliable”.

2.1 Validity of the approach

The American Psychological Association (1999) states that “validity refers to the degree to which evidence and theory support the interpretation of test scores entailed by proposed uses of tests”. This definition applies to a wide range of tests, such as tests constructed to measure depression as well as tests devised to measure academic achievement. Definitions of validity within the context of student assessment in Higher Education are largely available in the related literature. Miller et al. (1998), for instance, state that “a test is said to be valid when it measures the extent to which the objectives of the teaching programme have been achieved”. In a similar vein, Dunn et al. (2003) describe a valid assessment as one that is meaningful, useful, and measures “the performance of the intended learning outcomes specified”.

There are different types of validity (Miller et al., 1998), and the types that were considered to be of interest to this research are face validity, content validity and construct validity. These are discussed next.

Face validity

Miller et al. (1998) state that “an assessment task is said to have face validity if a number of judges – ranging from experts in the field to students – agree that the test item is valid”. Face validity is concerned with the extent to which, academic staff and students alike, agree that a test is a valid method to measure what it is intended to measure.

Reports from students in focus group studies, such as (Lilley & Barker, 2002) support the view that a test based on the CAT approach "looked valid" to them. Furthermore, Lunz et al. (1992) suggest that CATs where the review of previously entered responses is allowed, such as the CAT software prototype developed for this research (Lilley & Barker, 2004), are likely to have greater face validity than those CATs where review is not permitted. This is important in terms of face validity, since CBTs usually have this facility and tests that do not are likely to be considered less valid in a real context.

Findings from the academic staff evaluation reported in Lilley & Barker (2006), were taken to indicate that the CAT approach was valid in both formative and summative assessment settings, with a greater face degree of validity in the former. Although Miller et al. (1998) amongst others recognise the importance of face validity, doubts have been expressed about its rigour. Anastasi (1988), for instance, argues that face validity is “not validity in the technical sense” and proposes that other forms of validity testing, such as content validity, are required.

Content validity

Content validity is concerned with the extent to which the content of a test satisfactorily represents the subject domain (or syllabus) being assessed (American Psychological Association, 1999). One way to evaluate whether a test has sufficient content validity for a given purpose, would be the analysis, by subject domain experts, of the relationship between the test content and the intended learning outcomes. Hambleton & Rogers (1991) state that “expert judgement is the main mode of investigation of a test’s content validity”. Content validity is of particular importance in order to avoid the inclusion of irrelevant elements, the under-representation of core components, and the overemphasis of certain elements within the subject domain being tested.

Validity based on test content is often a laborious task in the context of CATs, as the recommended number of questions required in the question bank is, at least, 4 times the number of questions to be administered in a test sitting. It should be noted that questions should be evenly distributed across the different ability levels. Validity based on test content is a well established technique, and it is often part of the regular internal and external moderation processes in Higher Education institutions (Miller et al., 1998; Rhodes & Tallantyre, 2003).

The CAT approach, as implemented as part of this research, was based on the use of objective questions such as multiple-choice and multiple-response. Ward (1981) identified contributing factors that relate to the validity of objective tests in general, such as: “good syllabus coverage”, “consistent syllabus coverage from year to year”, “compulsory questions”, “results less influenced by irrelevant abilities” and “precise questions”. Such factors can also be applied to support the view that the CAT approach, as implemented in this work, has content validity.

2.2 Construct validity: empirical study.

Construct validity is “the measure of the underlying theory or construct of a particular test or examination” (Brown, 1997). Construct validity is concerned with the degree to which a test assesses the underlying theoretical construct it is intended to measure. In this research, construct validity is concerned with the extent to which CAT proficiency level estimates are interrelated to scores obtained by other traditional assessment methods intended to measure similar learning outcomes. To investigate the construct validity of the CAT approach, an empirical study was conducted in which a group of test-takers participated in three different assessment methods, namely computer-adaptive test, computer-based test and practical programming test. The questions employed in this study were analysed by two subject experts with the purpose of ensuring content validity.

Method. As part of their regular assessment for a programming module, a group of 125 Level 2 Computer Science undergraduates participated in three assessments. The assessments are summarised in Table 1. All assessments took place in computer laboratories, under supervised conditions.

Assessment	Brief description
Computer-based test (CBT)	Test-takers were asked to answer 10 predefined questions
Computer-adaptive test (CAT)	Test-takers were asked to answer 30 dynamically selected questions
Programming Test	Test-takers were asked to write a computer program using Visual Basic, based on an unseen program specification.

Table 1: Summary of assessments undertaken by participants

Summary of test-taker performance. Test-takers’ performance in three assessments is summarised in Table 2. In this table it can be seen that the possible scores for the CBT and practical programming test ranged from 0 (lowest) to 100 (highest). The possible scores for the computer-adaptive test ranged from -3 (lowest) to +3 (highest).

Assessment	Mean	Std. Dev.
Computer-based test	36.96	18.41
Computer-adaptive test	0.16	1.23
Practical programming test	44.52	25.38

Table 2: Summary of test-taker performance (N=125)

Findings. In order to investigate the correlations between CAT proficiency level estimates and other assessment methods intended to measure similar learning outcomes (i.e. CBT and programming test), the results shown in Table 2 were subjected to a Pearson's Product Moment correlation. This is shown in Table 3 below.

Assessment		Practical programming test	CBT
CAT	Pearson Correlation	0.428	0.548
	Sig. (2-tailed)	0.000	0.000
CBT	Pearson Correlation	0.221	*
Test	Sig. (2-tailed)	0.013	

Table 3: Pearson's Product Moment correlation results (N=125)

The significant correlation observed between the CAT and the practical programming test ($r=0.43$, $p<0.001$) and between the CAT and the CBT ($r=0.55$, $p<0.001$) are an important finding, and were taken to support the claim that the CAT approach has construct validity. The results shown in Table 3 show that those performing well on the CAT test also performed well on the other two test formats. The correlation between the CBT and the practical programming test, although significant was smaller than either correlation with the CAT ($r=0.22$, $p<0.01$). This supports the view that the test-takers were not disadvantaged by the CAT approach.

Up to this point, this chapter has focused on validity issues. However, a test that is valid is not necessarily reliable and vice-versa. Reliability issues were also of importance to this research, and the next section of this chapter is concerned with these issues.

2.3 Reliability of the approach

Reliability is "the degree to which test scores for a group of test-takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test-taker" (American Psychological Association, 1999). Ward (1981) adds that an assessment is reliable when "it applies a consistent standard of measurement

to all students and in all years". In general terms, one can argue that reliability refers to the extent to which assessments are consistent. On the topic of test reliability, Miller et al. (1998) warn that "it is unrealistic to expect to achieve 100 percent reliability" and that the aim should be to construct tests that are "as reliable as possible".

In a similar vein to test validity, there are factors that contribute towards test reliability that are generic to objective tests rather than exclusive to the CAT approach. These factors are explored next.

Contributing factors

Ward (1981) identified three factors that contribute to the reliability of objective tests. Two of these: "reliable marking" and "assessment of student's own work" are of relevance to this research. These factors are discussed next.

Reliable marking. In the implementation of the CAT approach employed for this research, all questions are marked consistently and objectively by the software application.

Assessment of student's own work. Ward (1981) argues that objective tests are often conducted under supervised conditions, and this can increase assessment reliability. The reason for this is that such a scenario would involve some form of authentication, and therefore it would be relatively straightforward to ensure that results obtained by test-takers were based solely on their own work.

The two factors above both contribute to reliability rather than measuring it. The next section of this chapter, discusses how one approach to measuring reliability, namely test-retest reliability, was applied to this work.

2.4 Test-retest reliability: empirical study

In a test-retest reliability study, the same group of participants are subjected to two different forms of the same test. The reliability is considered to be the correlation between the scores of both tests. In order to investigate the reliability of the CAT approach, an empirical study was performed as part of this work.

Method. A group of 133 Level 2 Computer Science undergraduates enrolled on a programming module took part in two sessions of summative assessment using the CAT software prototype developed for this research. The characteristics of these two sessions are summarised in Table 4.

The CAT software prototype developed for this research was modified to include a traditional computer-based test (CBT) component, in order to administer a predefined set of questions to all participants. Prior to the first session of assessment using the modified CAT software prototype, test-takers were given a brief introduction to the use of the software, but were not

informed of the existence of two sections within the test (i.e. CBT followed by CAT). In both sessions of assessment, the order in which the CBT questions were presented was randomly selected, as an attempt to minimise unauthorised collaboration amongst test-takers.

In addition to the two computer-delivered assessments, participants were required to undertake two additional assessments as part of their programming module. These two assessments are summarised in Table 4.

Assessment	**Brief description**
1. In-Class Test 1	10 predefined questions (i.e. CBT mode) followed by 10 questions dynamically selected (i.e. CAT mode).
2. In-Class Test 2	10 predefined questions (i.e. CBT mode) followed by 20 questions dynamically selected (i.e. CAT mode).
3. In-Class Programming Test	Test-takers were asked to write a computer program using Visual Basic, based on an unseen program specification.
4. Practical project	Participants were asked to produce a straightforward high fidelity software prototype, according to a brief, over a period of 4 weeks.

Table 4: Summary of assessment employed for the group of participants

With exception of the practical project (i.e. Assessment 4), all assessment sessions listed in Tables 4 & 5 were conducted under supervised conditions in computer laboratories.

Summary of test-taker performance. A summary of the test-takers' performance in each of the assessments is presented in Table 5 below.

Assessment		**Mean score**
Assessment 1	CBT 1	51.5%
	CAT 1 (proficiency level)	-0.832
Assessment 2	CBT 2	42.3%
	CAT 2 (proficiency level)	-0.909
Assessment 3		
In-Class Programming Test		49.7%
Assessment 4		
Practical project		71.7%

Table 5: Summary of test-taker performance (N=133)

In Table 5, the potential CAT scores ranged from -2 (lowest) to +2 (highest). The remaining scores ranged from 0% (lowest) to 100% (highest).

Findings. An Analysis of Variance (ANOVA) was performed on the data summarised in Table 5, in order to test the significance of any differences in the means. The results of this ANOVA are shown in Table 6 below.

Between groups	**Probability (p)**
CBT Assessment 1 and Assessment 2	0.001
CAT Assessment 1 and Assessment 2	0.607
Assessment 3 (Programming Test) and Assessment 4 (Coursework)	0.001

Table 6: ANOVA results relating to the data summarised in Table 5 (N=133)

The results presented in Table 6 show that there was a significant difference between the number of questions answered correctly in the CBT element of assessments 1 and 2 ($p=0.001$). However, there was no significant difference between the CAT levels obtained by test-takers in assessments 1 and 2 ($p>0.60$). This is an interesting result, especially in consideration of the finding that the mean CBT performances in assessment 1 and 2 were significantly different ($p<0.001$). These results were taken to indicate that the CAT level is a reliable measure of test-taker ability, and possibly a better and more consistent measure than a simple test score.

There was also a significant difference observed in the performance of students on the two off-computer assessments (assessments 3 and 4, $p=0.001$). In order to further understand the implications of these findings, a Pearson's Product Moment correlation was also performed on the data collected from the four assessments, and the results of this analysis are shown in Table 7 below.

		CAT 1	CBT 1	CAT 2	CBT 2	Programming test	Practical project
CAT 1	Pearson Correlation	*	.849(**)	.617(**)	.548(**)	.552(**)	.377(**)
	Sig. (2-tailed)		.000	.000	.000	.000	.000
CBT 1	Pearson Correlation	*	*	.552(**)	.467(**)	.445(**)	.300(**)
	Sig. (2-tailed)			.000	.000	.000	.000
CAT 2	Pearson Correlation	*	*	*	.816(**)	.571(**)	.407(**)
	Sig. (2-tailed)				.000	.000	.000
CBT 2	Pearson Correlation	*	*	*	*	.527(**)	.398(**)
	Sig. (2-tailed)					.000	.000
Programming test	Pearson Correlation	*	*	*	*	*	.528(**)
	Sig. (2-tailed)						.000

Table 7: Pearson's Moment Correlation results (N=133)
(**) Correlation is significant at the 0.01 level (2-tailed).

The results of the Pearson's test shown in Table 7 are taken to indicate that the scores obtained by participants This was interpreted as indicating that a score obtained by a participant in one assessment is a reasonable and fair predictor of performance in any other. It can also be seen that there is a high correlation between scores in the CBT and the CAT sections of assessments 1 and 2. On average, participants who performed well in the CBT sections also performed well in the CAT sections and vice versa ($p<0.001$). It was also found that the CAT proficiency levels achieved by the participants in assessment 1 were highly correlated with the CAT levels in assessment 2. This was taken to indicate that:

- the CAT test was a fair reflection of participants' ability in the assessment;
- the CAT assessment was at least as good an indicator of the ability of a test-taker as the CBT component of the prototype;
- no participant was disadvantaged by the CAT approach.

3.0 Conclusion

This paper presented a range of issues related to the validity and reliability of the CAT approach: face validity, content validity, construct validity and test-retest reliability. It was of relevance to this work to show that the CAT approach complies with these well-established standards since it is crucial to all stakeholders in the student assessment process that assessment methods are both valid and reliable. As part of this work, two empirical studies were carried out are and reported in this chapter. Both studies were performed in a real educational context, as recommended by Laurillard (1993) and Barker & Barker (2002). The findings from these two empirical studies provided evidence to support the claims that:

- the CAT approach is, at least, as fair and accurate as other traditional computer-assisted assessment methods in measuring a test-taker's proficiency level within a subject domain,
- test-takers are not disadvantaged by the CAT approach,
- the CAT approach is both valid and reliable.

Furthermore, it was shown that several factors that contribute to the validity and reliability of objective tests can also be applied to the CAT approach. There is an increasing body of research supporting the validity and reliability of the CAT approach; for instance, Segall (2001), Wolfe et al. (2001b) and Segall et al. (2001) report on the validity of the CAT approach. Other research, such as the work by Schoonman (1989) and Moreno & Segall (2001), report on the reliability of the approach. Such research, however, focuses mostly on the validity and reliability of the CAT approach when compared with traditional objective tests using a paper-and-pencil format. The studies reported here are a useful addition to this body of research since they examined test interrelations between CAT proficiency level estimates and scores obtained using other forms of computer-assisted assisted assessments, rather than paper-and-pencil tests.

4.0 References

American Psychological Association (1999) *Standards for Educational and Psychological Testing*. American Educational Research Association.

Anastasi, A. (1988) *Psychological testing*. New York: Macmillan.

Barker, T. & Barker, J. (2002) 'The evaluation of complex, intelligent, interactive, individualised human-computer interfaces: What do we mean by reliability and validity?', *Proceedings of the European Learning Styles Information Network Conference*, University of Ghent, June 2002.

Barker, T. & Lilley, M. (2006) 'Measuring staff attitude to an automated feedback system based on a Computer Adaptive Test', *Proceedings of Computer-Assisted Assessment 2006 Conference,* Loughborough University, July 2006.

Barker, T.; Lilley, M. & Britton, C. (2006) 'A student model based on computer adaptive testing to provide automated feedback: The calibration of questions', *Paper presented at the Association for Learning Technology (ALT) 2006*, Herriot-Watt University, September 4-7, 2006.

Brown, G. (1997) *Asssessing Student Learning in Higher Education*. Lodnon: Routledge Falmer.

Brusilovsky P (2004) 'Knowledge Tree: A Distributed Architecture for Adaptive E-Learning', *Proceedings of WWW 2004*, May 17-22, New York, New York, USA, pp. 104-113.

Dunn, L.; Morgan, C.; O'Reilly, M. & Parry, S. (2003) *The Student Assessment Handbook: New Directions in Traditional and Online Assessment*. Lodnon: Routledge Falmer.

Hambleton, R. K. & Rogers, H. J. (1991) Advances in criterion-referenced measurement, in R. K. Hambleton & J. C. Zaal (Eds.) (1991), *Advances in Educational and Psychological Testing: Theory and Applications* (Evaluation in Education & Human Services), Kluwer Academic Publishers

Laurillard, D. M. (1993) *Rethinking University Teaching: A Framework for the Effective Use of Educational Technology*. Routledge, London.

Lilley, M. & Barker, T. (2002) 'The Development and Evaluation of a Computer-Adaptive Testing Application for English Language', *Proceedings of the 6th Computer-Assisted Assessment Conference*, Loughborough University, United Kingdom, pp. 169-184.

Lilley, M. & Barker, T. (2005) 'An empirical study into the effect of question review in a computer-adaptive test', *Proceedings of the 6th Annual Higher Education Academy Subject Network for Information Computer Science Conference*, University of York, United Kingdom.

Lilley, M. & Barker, T. (2004) 'A Computer-Adaptive Test that facilitates the modification of previously entered responses: An empirical study', *Lecture Notes in Computer Science*, 3220, 7th International Conference ITS 2004, Volume 3220/2004, pp. 22-33, 2004.

Lilley, M. & Barker, T. (2006) 'Student attitude to adaptive testing', *Proceedings of HCI 2006 Conference*, Queen Mary, University of London, 11-15, September 2006.

Lilley, M.; Barker, T., Bennett, S. & Britton, C. (2002) ,How computers can adapt to knowledge: A comparison of computer-based and computer-adaptive testing', *Proceedings of the 1st International Conference on Information and Communication Technologies in Education*, Junta de Extremadura Consejería de Educación, Ciencia y Tecnología, Badajoz, Spain.

Lilley, M.; Barker, T. & Britton, C. (2004) 'The development and evaluation of a software prototype for computer adaptive testing', *Computers & Education Journal* 43(1-2), pp. 109-123.

Lord, F. M. (1980) *Applications of Item Response Theory to Practical Testing*. Lawrence Erlbaum Associates Inc.

Lunz, M. E.; Bergstrom, B. E. & Wright, B. D. (1992) 'The Effect of Review on Student Ability and Test Efficiency for Computerized Adaptive Tests', *Applied Psychological Measurement*,16(1), March 1992, pp. 33-40.

Miller, A.; Imrie, B.W. & Cox, K. (1998) *Student Assessment in Higher Education: A Handbook for Assessing Performance*. London: Routledge Falmer.

Moreno, K. E. & Segall, D. O. (2001) Validation of the Experimental CAT-ASVAB System, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Rhodes, G. & Tallantyre, F. (2003) Assessment of Key Skills in S. Brown & A. Glasner (Eds.) (2003), *Assessment Matters in Higher Education: Choosing and Using Diverse Approaches*. Society for Research into Higher Education, Open University Press.

Schoonman, W. (1989) *Applied Study on Computerized Adaptive Testing*. Swets & Zeitlinger.

Segall, D. O. (2001) The Psychometric Comparability of Computer Hardware, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing: From inquiry to operation*, Washington DC: American Psychological Association.

Segall, D. O., Moreno, K. E., Kieckhaefer, W. F., Vicino, F. L. & McBride, J. R. (2001) Validation of the Experimental CAT-ASVAB System, in W. A. Sands, B. K. Waters, & J. R. McBride (Eds.) (2001), *Computerized adaptive testing:*

From inquiry to operation, Washington DC: American Psychological Association.

Wainer, H. (2000) Introduction and History, in H. Wainer (2000), *Computerized Adaptive Testing: A Primer*, Lawrence Erlbaum Associates Inc.

Ward, C. (1981) *Preparing and Using Objective Questions*. Handbooks for Further Education. Nelson Thornes Ltd.

Wolfe, J. H.; McBride, J. R. & Sympson, J. B. (2001b) Development of the Experimental CAT-ASVAB System In W. A. Sands, B. K. Waters, & J. R. McBride (Eds.), Computerized adaptive testing: From inquiry to operation. Washington DC: American Psychological Association.

Acknowledgments

The research reported in this paper was undertaken in conjunction with Dr Mariana Lilley, Department of Computer Science, University of Hertfordshire.