Educational and Child Psychology
(Special Issue on 'Information Technology and Children's Learning')

(Vol. 18, No. 3, 2001, pp. 58-74)

Computer-based assessment in education

Chris Singleton

Department of Psychology
University of Hull

Chris Singleton is Senior Lecturer in Educational Psychology at the University of Hull, and a member of the British Psychological Society's Steering Committee on Test Standards. He co-editor of the book 'Psychological Assessment of Reading' (Routledge, 1997).

Address for correspondence: Dr Chris Singleton, Department of Psychology, University of Hull, Hull HU6 7RX, UK. Email: c.singleton@psy.hull.ac.uk

ABSTRACT

Previous studies have indicated that computerised assessment enables psychologists and teachers to carry out assessment of cognitive abilities and basic skills speedily and easily, and pupils generally have positive attitudes towards this type of assessment. To examine this view, 90 children aged 6–7 years were administered computer-delivered assessments of verbal ability (verbal concepts) and nonverbal ability (mental rotation), and the scores were compared with measures of verbal and nonverbal ability derived from conventional assessment using the British Abilities Scales (Second Edition). The results revealed an expected pattern of intercorrelations indicating that the different assessment formats did not significantly affect the ability being assessed. Contrary to some suggestions that computer assessment may favour boys, no gender differences were found in either the conventional or computer assessment formats. Most of the children expressed a preference for the computer format. These findings are evaluated in relation to the various advantages and disadvantages of computer-based assessment in education, and particular the dangers that because an assessment is computerised, users may overrate its psychometric integrity and hence its value. (175 words)

INTRODUCTION

Computer-based assessment (CBA) may be considered to include '…any psychological assessment that involves the use of digital technology to collect, process and report the results of that assessment' (BPS, 1999). The four components of CBA may be identified as (1) assessment generation, (2) assessment delivery, (3) assessment scoring and interpretation, and (4) storage, retrieval and transmission. The earliest widespread applications of CBA in education were for scoring student test forms (McCollough and Wenck, 1984). In a review ten years later, the National Council for Educational Technology (now the British Educational Communications and Technology Agency, BECTa) found that by far the greatest use of CBA in schools and colleges in the UK was still to record or report students' results, rather than for generation or delivery of assessment (NCET, 1994). However, there were some developments in 'paper-less' examinations (see Bull, 1994; Stephens, 1994). During that period, developments in the use of CBA in the USA were rather more advanced. CBA had become the standard for US college placement, had been incorporated within integrated learning systems to monitor students' progress, and had been applied to a number of problems in special education. The latter included regular curriculum-based assessment (which otherwise would be too labour-intensive), recording and analysing observational data (e.g. regarding a student's response to teaching), and development of expert systems in order to determine elegibility for special educational services (see Greenwood and Rieth, 1994; Woodward and Rieth, 1997).

Two key applications of CBA that were first developed in the USA have been tried and evaluated in schools and colleges in the UK. Computerised placement tests for college entry, such as *Accuplacer* (College Board, 1995), generally incorporate tests of reading comprehension, English proficiency, and maths. The items, which are mostly text-based, are selected from a large item bank, and are delivered adaptively, i.e. the difficulty of the items varies as a function of the student's success and failure. In 1995, 22 colleges in the UK participated in trails of Accuplacer (NCET, 1995). The results were sufficiently encouraging for the Further Education Development Agency (FEDA) to encourage further research into the development of computerised college entry assessments for use in the UK. Integrated Learning Systems (ILS), such as *Successmaker*, are essentially drill-and-practice programs networked for supporting the teaching of English and maths with large groups of students. A number of evaluations of ILS have been carried out in UK schools, with somewhat mixed results (NCET, 1996; Underwood et al, 1994; 1996). The success of such programs appears to depend largely on the extent to which teachers have fully integrated their use into the curriculum, and the degree of congruity between the teachers' educational objectives and those of the ILS (see Miller, DeJean and Miller, 2000; Underwood, 2000; Wood, Underwood and Avis, 1999). Any ILS must involve CBA

because the progress of each student has to be monitored by the system (just as any good teacher has to assess pupils — formally or informally — in order to decide whether they are ready to move on in the curriculum). However, as yet the CBA features of ILS have not been evaluated independently of the didactic ones and it remains to be seen whether computer programs are as good as teachers in evaluating the curriculum progress of students.

Equivalence of CBA with conventional test forms is a key issue that must be addressed if CBA is to become widely accepted in education (Singleton, 1997). Teachers and psychologists want to be assured that the use of CBA will yield data which matches that obtained by conventional testing in both validity and reliability. A similar problem has been confronted in the use of CBA in occupational psychology (Bartram, 1994). In the relatively early stages of the history of CBA, equivalence was an concern that arose largely in the translation of *existing tests* to CBA format. The issue of equivalence is particularly important where speed of response is critical to the test, because clicking with a mouse is quite different from, say, ticking a box on a page, or producing an oral response. In addition, reading text on a computer screen may be more difficult than conventional reading — in a review of the scientific literature, Dillon (1992) reported that reading from the computer screen has been found to be 20–30% slower than reading the paper-based text. Generally speaking, demonstrating equivalence of conventional assessments and CBA requires evidence of high correlation between the two formats.

Problems in establishing equivalence have been encountered in translating some tests to computer form (e.g. Dimock and Cormier, 1991; French and Beaumont, 1990), particularly where conventional tests have involved physical manipulation of items, e.g. in a block design test (Martin and Wilcox, 1989). However, many successes have also been reported. Maguire et al (1991) showed that a computerised version of the *Peabody Picture Vocabulary Test* produced equivalent responses to a conventional version. Wilson, Thompson and Wylie (1982) reported a significant correlation between conventional and computerised versions of the *Mill Hill Vocabulary Test*, and French and Beaumont (1990) found a correlation of 0.84 between a computerised version of *Ravens Matrices* and the conventional version, with the computerised version taking less time to complete. These authors took the view that, as higher resolution graphics become more widely available, equivalence between conventional and computerised versions of visual tests such as Raven's Matrices would improve further. Evans, Tannehill and Martin (1995) obtained high, significant correlations between conventional and computerised versions of the revised form of the *Woodcock-Johnson Psycho-Educational Battery* Tests of Achievement.

More recently we have seen the development of tests that have been designed *specifically* for the computer. Since these have no conventional counterparts, substantiating 'equivalence' is a matter of validation (concurrent and/or predictive) against established tests that measure the same or similar

characteristics or against objective criteria. Indeed, in some of the newer examples of CBA it would be impossible or impractical to carry out the same assessment by conventional methods, because the tests use animated graphics and sound, as well as recording item response times to the level of milliseconds.

Singleton, Horne and Vincent (1995) reported significant correlations, in the region of 0.6, between a pilot version of a computerised reading comprehension test and scores on the *Edinburgh Reading Test* (Godfrey Thompson Unit, 1993) for 75 children in Years 3–5. In general, the children (especially those of lower ability) expressed a preference for the CBA over the pen-and-paper assessment of reading. Cisero, Royer, Marchant and Jackson (1997) reported on a CBA designed to diagnose specific reading disability in college students, called CAAS (Computer-based Academic Assessment System). This system was based on comparisons between measures of word reading and phonological processing (at which students with specific reading disability were poor) and measures of category matching and semantic knowledge (at which students with specific reading disability were not poor, but which students with more general learning difficulties were poor). The criterion was to differentiate students classified by college disability services as having learning problems, from nondisabled students. However, this program is not fully computerised as several of the tasks (e.g. word naming) require an examiner to listen to the student and press a 'correct' or 'incorrect' button to record their response.

CoPS (Cognitive Profiling System) was a pioneering development specifically for computer delivery, designed to provide early identification of children at risk of literacy difficulties (Singleton, Thomas and Leedale, 1996). Like the CASS system reported by Cisero et al (1997), CoPS measures response times as well as accuracy, but unlike CAAS, CoPS is also fully computerised, all responses being made by the child with the mouse or touch screen. The program comprises eight tests of cognitive abilities that have been shown to be precursors of literacy difficulties, including phonological awareness, phoneme discrimination and auditory and visual short-term memory. The tests in CoPS, which in the form of games and invariably enjoyed by children, have been standardised and normed for the 4–8 years age range. CoPS was validated both by concurrent correlation with conventional cognitive measures, and also by successful prediction of literacy difficulties some three years after the original CoPS assessment at age 5, using multiple regression and discriminant function analysis (Singleton, Thomas and Horne, 2000). 421 children were assessed with CoPS at age 5 and follow-up assessments using conventional tests of reading and general ability were carried out at 6 and 8 years of age. Correlations between the CoPS tests administered at age 5 and reading ability at age 8 were in the region of 0.6 for auditory-verbal memory and phonological awareness, and in the region of 0.3 for the *CoPS* measure of auditory discrimination as well as most of the other memory measures. Linear regression analyses showed that the CoPS tests of auditory-verbal memory and

phonological awareness administered at age 5 together accounted for 50% of the variance in reading ability at age 8, compared with only 29% of the variance being attributable to intelligence. Discriminant function analysis showed that CoPS tests predicted poor reading skills to acceptable accuracy, with very low or zero rates for false positives and false negatives. By contrast, a word recognition test given at age 6 was not found to predict reading at age 8 to the same degree of accuracy, resulting in a false positive rate of 21%. Measures of verbal and nonverbal ability at age 6 produced high false positive rates between 50% and 70%. CoPS is now used in over 3,500 schools in the UK and elsewhere in the world and two foreign language versions have been developed (Swedish and Italian). The success of CoPS has led to the creation of similar systems for older students, including *LASS Secondary* (Horne, Singleton and Thomas, 1999), *LASS Junior* (Thomas, Singleton and Horne, 2001) and *LADS* (Singleton, Thomas, Leedale and Horne, in press). The last of these is a program for dyslexia screening in the 16+ age range, comprising tests of working memory and phonological processing.

CoPS Baseline Assessment is a normative, standardised CBA designed specifically for computer delivery, and is accredited by the Qualifications and Curriculum Authority for school entry assessment at age 4–5½ (Singleton, Thomas and Horne, 1998). CoPS Baseline comprises four modules: (1) literacy, (2) mathematics, (3) communication and (4) personal and social development. The first two modules are adaptive, fully computerised, psychometric tests in which the child responds using the mouse or touch screen. The third is a computer-delivered animated story that provides a basis upon which the teacher can score the child's expressive language and enter this into the computer, and the fourth module is a teacher rating scale. As well as automatically scoring and analysing all results the computer also produces automatic printed reports for teachers and parents (QCA regulations require that the outcomes of baseline assessment must be shared with parents). This program has also been predictively validated against literacy and numeracy development one and two years later (Singleton, Horne and Thomas, 1999; Singleton and Horne, in press; Singleton et al, in preparation). 153 children were assessed on CoPS Baseline at an average age of 4 years 10 months, and the data indicated that the computerised baseline assessment module produced a satisfactory distribution of scores across the intended age range, and the shorter adaptive forms of the literacy and maths tests correlated highly with the full versions ($r \approx 0.8$). When progress in reading and mathematics was followed up it was found that baseline scores from the computer tests gave a good overall prediction of reading (*British Ability Scales Word Reading Test* — Revised Edition) and maths (*Wechsler Objective Numeracy Dimensions*) over the first two year of schooling ($r \approx 0.7$). Correlations between the 8 skill/concept areas that comprise the baseline assessment and later ability in reading and maths were consistent with other findings reported in the literature. However, discriminant function analysis indicated that the baseline tests were not a particularly good screening

device for predicting learning difficulties, since there were high levels of false positives and false negatives. This is not a failing of CBA *per se*, but rather a characteristic of all QCA accredited baseline systems, since they have been designed to conform to the *National Framework for Baseline Assessment* (SCAA, 1997), which precludes assessment which is sufficiently detailed to provide early identification of special needs. Since the publication of CoPS Baseline, one other QCA accredited baseline system (PIPS) has produced a computer-delivered version of what was previously only a conventionally administered assessment scheme (see Tymms, 1999).

The present study was designed to address some of the issues raised above in the context of computerised assessment of verbal and nonverbal abilities — focusing, in particular, on 'equivalence' between CBA and conventional forms of assessment. Measuring students' verbal and nonverbal abilities is seen as a basic requirement in schooling, both for the purposes of monitoring overall progress and for assessing special needs. Many schools routinely assess children's verbal and nonverbal abilities by means of group tests such as the *Cognitive Abilities Test* (Thorndike, Hagen and France, 1989) or the *NFER-NELSON Verbal and Nonverbal Reasoning Tests* (Hagues and Courtnay, 1993; Smith and Hagues, 1993). Individually-administered tests of verbal abilities, such as the *British Picture Vocabulary Scale* (Dunn, Dunn, Whetton and Burley, 1997), and nonverbal abilities (e.g. *Matrix Analogies Test*; Naglieri, 1985) are used by many teachers (especially those in the special needs field), while psychologists employ specialised tests such as WISC-III and BAS–II, which incorporate both verbal and nonverbal assessment. In such an important area, the addition of a computer-delivered assessment would be a significant development.

The computer tasks used in this study (Verbal Concepts and Mental Rotation) were not designed to replicate exactly the conventional assessments employed, which were taken from the *British Ability Scales* (Second Edition). Rather, the intention was to explore the potential of CBA by using computerised tasks within the same broad cognitive domains, which would be stimulating and challenging to children, as well as maintaining their interest. Although a degree of equivalence between conventional and computer forms was predicted, because the two forms did not measure *exactly the same* abilities, a high level of equivalence (in terms of large correlation values) was not to be expected. Of equal importance to this study was the pattern of relationships between the different measures and the responses of the children to them. A further aspect of interest in this study was that of possible gender differences. Some studies have reported that girls tend to show less interest than boys do in computer activities (e.g. Crook, 1994; Durndell, 1991; Hughes, Brackenridge and Macleod, 1987; Scott, Cole and Engel, 1992). However, most of these studies were carried out on children older than those in the present study (6–7 years). At this age, Crook and Steele (1987) observed no significant gender differences amongst pupils voluntarily choosing to engage in computer

activities in the classroom. Nevertheless, if gender differences in interest in computers do exist at this age, then clearly it is possible that these could bias the results of a CBA.

METHOD

Participants

The participants in this study were 90 children (49 boys, 41 girls) aged 6–7 years attending four primary schools in Hull, a city in the north of England. The mean age of the sample was 6 years 6 months (S.D. 4.1 months). The children were selected at random from the class registers.

Materials

(a) Computer-based assessments

Two untimed experimental CBAs were used in this project: one was designed to assess verbal ability and the other nonverbal ability. Both were created by Singleton, Thomas, Leedale and Horne and have not been published, although *CoPS Baseline* (Singleton, Thomas and Horne, 1998; Singleton, Horne and Thomas, 1999) incorporates some elements of the verbal ability test.

(i) *Verbal Concepts.* This is a test of 40 items, preceded by two practice items. In each item five pictures are presented on the monitor screen and the computer gives spoken instructions asking the child to identify the picture that is associated with a verbal concept, e.g. "Which picture goes best with the word 'cooking?" (The pictures were: paintbrush, bed, hammer, vacuum cleaner, cooker.) The child responds by clicking on the chosen picture using the mouse. Spoken feedback is given by the computer in the practice items only, e.g. "No, that picture does not go best with the word — . Try again." The child attempts all items and the test score is the number of items correct out of 40.

(ii) *Mental Rotation.* This is a test of 24 items, preceded by four practice items. The task, which involved a rotund character called 'Zoid', requires mental rotation. In each item, Zoid appears in the centre of the monitor screen, accompanied by certain accessories, e.g. wearing boots and carrying a satchel. Then, four of Zoid's 'friends' (all of whom had the same body shape) are shown, each in a different rotated orientation, also carrying various accessories (see Figure 2). Rotation occurs in either in the horizontal plane (e.g. upside-down) or vertical plane (i.e. back-to-front), or both, and items are ordered in increasing difficulty. The scenario presented to the child is that Zoid is on holiday, and that his friends are attempting to copy Zoid. The computer gives spoken instructions to the child to identify the friend that is copying Zoid correctly. The child responds by clicking on the chosen friend using the mouse. The

practice items incorporate animation in which the chosen friend rotates on the screen so that the child can both understand the nature of the task and verify the accuracy of their choice. Spoken feedback is given by the computer in the practice items only, e.g. “No, that friend is not copying Zoid correctly. Try again.” The child attempts all items and the test score is the number of items correct out of 24.

(b) Conventional assessments

Four untimed conventional tests were used in this study; these were taken from the *British Ability Scales*, Second Edition (Elliott, Smith and McCulloch, 1996) and each incorporates starting rules according to age and discontinuation rules according to performance (for details see the test manual). The tests were scored according to the procedures set down in the test manual, and the results comprised ability scores (which are not adjusted for age) and T scores (age adjusted).

(i) *Word Definitions.* This test is designed to measure knowledge of word meanings by means of oral expression. The child is required to define words, e.g. scissors.

(ii) *Verbal Similarities.* This test is designed to measure verbal reasoning. The child is told three stimulus words and asked to explain how these go together, e.g. “Banana, apple orange” (A: They are all fruits.)

(iii) *Matrices.* This nonverbal test is designed to measure inductive reasoning through the application of rules governing relationships among abstract figures. The child is shown a matrix of abstract figures and has to select from among six choices the figure that correctly completes the matrix.

(iv) *Quantitative Reasoning.* This nonverbal test is designed to measure inductive reasoning through the application of rules concerning sequential and numerical patterns. The child is shown a set of dominos and has to complete the pattern by drawing.

The scores from these BAS-II tests were combined to generate a Verbal Ability standard score (Word Definitions plus Verbal Similarities) and a Nonverbal Ability standard score (Matrices plus Quantitative Reasoning).

Procedure

Children were tested individually in a quiet area of the school near to their classroom. Testing was carried out in three sessions on separate days in order to avoid excessive fatigue; two sessions each comprising one verbal and one nonverbal test from BAS-II, and one session comprising the two computerised tests. The order of sessions was randomised. After all the assessment sessions had been completed, children were asked which type of assessment they preferred, and why.

RESULTS

Table 1 shows the means and standard deviations of scores obtained in both the computerised and conventional tests. Since T scores are distributed with a population mean of 50 and standard deviation of 10, it can be seen that the sample fell well within the expected range on all four BAS–II tests. Only Verbal Similarities showed a slightly restricted variance. Similarly, the standard scores for BAS–II Verbal and Nonverbal Ability are both in the 'good average' range, although the variance for Nonverbal Ability is a little on the large side. Overall, however, these data indicate that the sample was not in the least exceptional in intellectual or ability terms.

Table 1 about here

As the CBA tests used in this study have not been standardised, direct comparisons between the BAS–II means and the means for the Verbal Concepts and Mental Rotation tests cannot be made. However, scores obtained on both Verbal Concepts and Mental Rotation approximated to a normal distribution, which suggests that these tests adequately tap abilities across the expected range. It should be noted that Verbal Concepts had a slightly restricted range, which is also signalled by the lower standard deviation than might have been expected (see Table 1). This corresponds with a similarly restricted range for BAS–II Verbal Similarities, suggesting that this was a particular feature of this sample, rather than a characteristic or flaw in either test.

Table 2 about here

Table 2, which shows Pearson Product Moment correlation coefficients between the CBAs and the BAS-II verbal and nonverbal ability standard scores, reveals an expected pattern of relationships between the variables. The verbal measures [CBA Verbal Concepts and BAS–II Verbal Ability] yield the highest correlation ($r = 0.51$), followed by the nonverbal measures [CBA Mental Rotation and BAS–II Nonverbal Ability] ($r = 0.42$). Although statistically significant, these correlation coefficients are not particularly high. Nevertheless, they do indicates that the related measures are tapping broadly the same domains of ability, despite being administered in quite different ways, one by computer and the other by a human assessor.

Table 3 about here

Table 3 provides a further breakdown of relationships between the CBAs and the BAS–II tests, which confirms the conclusions drawn above. Mental Rotation is significantly correlated with both Matrices and Quantitative Reasoning, but not with either Word Definitions or Verbal Similarities. On the other hand, Verbal Concepts is significantly correlated with all four BAS–II tests, albeit a higher levels for the verbal tests than for the nonverbal tests.

Table 4 about here

To evaluate possible gender differences in the results, a one-way analysis of variance was carried out, comparing boys' and girls' scores for the CBAs and the BAS–II verbal and nonverbal ability measures (see Table 4). Although the girls obtained slightly higher mean scores on all except the CBA Verbal Concepts test, in fact no significant gender differences were found on any of the tests (CBA Verbal Concepts: $F = 1.06$, $p = 0.31$; CBA Mental Rotation: $F = 1.83$, $p = 0.18$; BAS–II Verbal Ability: $F = 1.22$, $p = 0.27$; BAS–II Nonverbal Ability: $F = 1.39$, $p = 0.24$). Hence there is no evidence that gender differences affected any of these tests.

Table 5 about here

Finally, the children were asked which type of assessment they liked best and why. The results are shown in Table 5. As some of the cells have a frequency of less than 5, these data are not susceptible to analysis by chi-square. However, it is contended that although there is a slight trend for boys to prefer the CBA more than the girls do (92% of the boys preferred CBA, compared with 78% of the girls) the difference is unlikely to be significant. Overall, CBA gets a massive popularity endorsement, with 86% of the children preferring it to conventional assessment. Many reasons were expressed for this, the chief being that CBA was "more fun".

DISCUSSION

The main findings of this study may be summarised as follows. Firstly, scores obtained on two CBAs measuring different cognitive abilities correlated significantly with comparable measures taken from an established psychometric test (BAS–II) that was administered conventionally. Secondly, children expressed an overwhelming preference for the CBA over the conventional tests, the former being perceived as more enjoyable. Thirdly, no significant gender differences in scores were found on any of the CBAs or conventional tests, nor in preferences for one or other type of test. Taken together, these findings suggest that computer-based systems have considerable potential in the field of educational assessment, particularly if the tasks can be made enjoyable for children. However, there are some important reservations. First, the correlations between the CBAs and the conventional tests were not exceptionally high (the key values were in the region of 0.4 – 0.5). This may be partly accounted for by the fact that the tests employed were not measuring *exactly* the same cognitive skills. But it is also possible that the elimination of verbal responses by the CBA precludes an important component of cognitive assessment, i.e. the ability to use language to express the products of thought. On the other hand, a higher correlation was obtained between CBA Verbal Concepts and BAS–II Verbal Ability ($r = 0.51$) than between CBA Mental Rotation and BAS–II Nonverbal Ability ($r =$

0.42), despite the fact that no expressive language skills are required in either CBA Mental Rotation or BAS–II Nonverbal Ability. It may also be observed that the correlation between the two CBAs (r = 0.33) is somewhat lower than that between the BAS-II Verbal Ability and BAS–II Nonverbal Ability (r = 0.40), suggesting that the CBAs measure cognitive skills that have less in common with each other. It is interesting, however, that there is a low (but still significant) correlation between the CBA Verbal Concepts score and the BAS–II Nonverbal Ability score (r = 0.30), but the converse does not hold — i.e. the relationship between the CBA Mental Rotation score and the BAS–II Verbal Ability score is small and not statistically significant. Interpretation of this finding is open to debate, but it is possible that the three measures CBA Verbal Concepts, BAS–II Verbal Ability and BAS–II Nonverbal Ability all have higher g (general intelligence) loadings, while CBA Mental Rotation is a more specific ability with somewhat lower g loading.

Although the principal reason given for preferring CBA to conventional assessment was that it was “more fun”, other pertinent reasons were given. Many children said they preferred CBA because they “liked playing on the computer”, and that it was “different to normal work”. Several mentioned that they liked the “funny pictures” in the CBA, and some specifically said they liked the character Zoid, because “he looks funny”. A few commented that they did not like the conventional assessments because they “took a long time to do”, which was surprising in view of the BAS-II discontinuation rules, designed to avoid tests being unnecessarily long. As test duration was not measured in this study it is not possible to tell whether the conventional tests did actually take longer to administer; perhaps children were influenced by the fact that there were four conventional tests and only two CBAs, so the overall time devoted to conventional assessment was certainly much greater. Of those children that expressed a preference for the conventional tests, the principal reason given was that they enjoyed talking to the assessor. Overall, it was concluded that CBA was greatly preferred, but that this finding may have been distorted by the imbalance of time spent and the greater number of conventional tests used in the study.

Based upon the findings of this study as well as conclusions in the literature as a whole, the principal advantages of CBA over conventional assessment in education may be tentatively summarised as follows.

1. **Greater precision in timing and delivery of test items, and measurement of responses** — this is particularly important where timing is critical, e.g. in assessment of short-term memory (Singleton and Thomas, 1994) but less so in the present study. Nevertheless, had they been programmed to do so, the CBAs employed in this study could have measured response times. Response time data would have augmented the value of the assessment, enabling various important distinctions to be drawn, e.g. between children who are accurate *and* fast, and those who are accurate but much slower in their responses (e.g. Singleton, Thomas and Leedale, 1996).

2. **Standardised presentation** — the test is *exactly the same* for all recipients, whereas with human administration some variation is inevitable. Arguably, this helps to improve reliability of measurement (Singleton, 1997).

3. **Savings in labour, time and cost** — since the computer does most of the work of assessment, including administering items, recording responses and scoring results, labour and cost savings in using CBA compared with conventional assessments can be significant (French, 1986; Woodward and Rieth, 1997). Several studies have shown that in comparisons of conventional and computerised versions of tests, teachers prefer the latter, mainly because results are immediately available, saving time in scoring responses and calculating standard scores (Fuchs, Fuchs and Hamlett; 1993; Fuch et al, 1987; Wesson et al, 1986; Woodward and Rieth, 1997). However, since teachers were not involved in administering the tests in the present study, their views were not solicited. Time savings are particularly marked when assessment is adaptive, i.e. where the difficulty of items selected from a test bank is varied in response to the student's progress on the test. The term 'adaptive testing' refers to any technique that modifies the nature of the test in response to the performance of the test taker, although typically it is used in connection with tests that are developed by means of Item Response Theory (Hambleton and Swaminathan, 1985). Conventional tests are *static* instruments, fixed in their item content, item order, and duration. By contrast, computer-based assessment can be *dynamic*. Since the computer can score performance at the same time as item presentation, it can modify the test accordingly, tailoring it more precisely to the capabilities of the individual taking the test. In conventional tests, for some part of the time, the individual's abilities are not being assessed with any great precision because the items are either too difficult or too easy (which can easily lead to frustration and/or boredom). In a computerised adaptive test, however, because the program contains information about the difficulty level of every item in the item bank (based on pass rates in the standardisation population) the individual taking the test can be moved swiftly to that zone of the test that will most efficiently discriminate his or her capabilities. This makes the whole process speedier, more reliable, more efficient, and often more acceptable to the person being tested. It has been shown that an adaptive CBA can take only a quarter of the time taken to administer an equivalent conventional test (Olsen, 1990).

4. **Increased motivation** — students with special educational needs have been found to display negative responses to conventional tests given in pen-and-paper format or administered by a human assessor (Wade and Moore, 1993). By contrast, many studies have reported that children and adults, particularly those with low ability or who have 'failed' educationally, feel less threatened by CBA than by conventional assessment, and hence prefer CBA (Moore, Summer and Bloor, 1984; French, 1986; Singleton, 1997; Singleton, Horne and Vincent, 1995; Skinner and

Allen, 1983; Watkins and Kush, 1988). In an adaptive test, students are not exposed to items that would be ridiculously easy nor impossibly difficult for them, which also enhances test motivation (Singleton, 1997). In the present study, children displayed a clear preference for the CBA over the conventional assessment.

5. **Development of innovative types of assessment** — e.g. Singleton, Thomas and Leedale (1996) pioneered the use of a computerised 'game' format in CoPS to assess very young children who would otherwise have been difficult to assess by conventional means. Use of animation, sound and colour can help to sustain a child's interest in the task. Assessment parameters that would be impractical to measure conventionally (such as item response times, as in CoPS) can easily be incorporated into the system and normed to give the assessor additional information about a child's performance. Judging from their comments, the CBAs used in the present study were regarded by the children as being more like 'games' than 'work'. This appears to have been a significant motivator and the principal reason for the preference for the CBAs over the conventional tests. However, it should be emphasised that some of the items in the CBAs were extremely challenging, e.g. one item in Verbal Concepts was "Which picture goes best with the word 'magnify'?" (The pictures were: television, fly, telescope, magnet, pen). Quite a few of the children spontaneously commented that the items in Mental Rotation were hard, especially those that involved simultaneous rotation in both planes. Thus it is certainly not the case that the CBAs were liked because they were easier.

6. **Assessment in special education —** most notably, in assessment of children with severe physical disabilities or profound sensory impairments. However, a wide variety of valuable applications of CBA in special education have been noted (for review see Woodward and Rieth, 1997).

7. **Wider access to up-to-date assessments** — in particular, the potential (yet to be realised) of the internet for delivery of the most up-to-date version of any specified CBA to any classroom in the world on demand. Arguably, this addresses a common problem in education, namely the use of old or obsolete tests that have outdated norms. However, the issues of national standardisation, as well as linguistic and cultural appropriateness, remain significant challenges for a truly international application of CBA in education.

Not surprisingly, however, there remain some important disadvantages and risks of CBA:

1. **Development time and associated costs** — due to programming and other technical requirements, the creation, validation and standardisation of any test in computer form generally takes longer and is more expensive to develop than an equivalent conventional form. This means that CBAs are likely to be more expensive for schools and psychologists to purchase initially,

although this expense is likely to be offset by savings in personnel time and costs of administration.

2. **Limitations in certain aspects of assessment** — CBA is best suited to measurement of cognitive and intellectual abilities, fact knowledge, basic skills and curriculum content. That leaves several important aspects of behaviour that are currently impossible (or impractical) to measure using CBA, including expressive language, social and emotional behaviour, and probably any assessment where large amounts of text reading are required.

3. **Risks of technology failure** — although these can never be eliminated, experience suggests that hardware and software reliability is steadily increasing.

4. **Risks of abuse** — Bartram (1994) acknowledged that CBA presents '…new problems for professional practice and ethical standards', and the British Psychological Society recognised that '…the ease of CBA construction does pose a serious threat to effective and fair practice in the use of psychological tests' (BPS, 1999). Because of its professional appearance, a CBA may have a spurious appearance of objectivity (and even infallibility) but may not necessarily conform to accepted standards of test construction and validation. The ease of use of CBAs, although one of their principal advantages, also creates potential dangers because users who do not properly understand the nature of psychological or educational assessment may be tempted to use them and may misinterpret findings as a result. For this reason, the BPS formulated guidelines for the development and use of CBA (BPS, 1999). Central to these guidelines is the view that all CBAs should be supported by clear documentation on rationale, validity, appropriateness, administration and interpretation, and that users need to be aware of what constitutes best practice in CBA, so that they can make informed evaluations and choices between available systems.

On balance, it is argued that the advantages and potential benefits of CBA in educational assessment and educational psychology far outweigh their disadvantages and potential risks. Like conventional assessments, the data derived from CBAs is valuable in monitoring progress of all pupils, in general school management, as well as in identifying and assessing children with special educational needs. However, CBAs have the particular benefit of enabling teachers to carry out assessments that otherwise would require rather large amounts of time in learning test administration procedures, and in delivering and scoring tests. Furthermore, CBAs can enable young children (and disaffected pupils) to be assessed in a way that is motivating and enjoyable for them, which is a benefit for educational psychologists as well as teachers. The most significant disadvantage of CBAs is clearly the risk of abuse, but that is also present in conventional assessment where untrained or inexperienced users are concerned. It is therefore timely that the British Psychological Society is currently in the process of instigating a programme of training and accreditation for personnel in

education in the use of psychometric tests, comparable with basic training and accreditation in the use of occupational tests, which has been available for some years (BPS, 1995). It is to be hoped that this initiative, together with the widespread adoption of guidelines on the development and use of CBAs (BPS, 1999) will enable teachers to take full advantage of CBAs and use them confidently and with professional acumen.

REFERENCES

Bartram, D. (1994) Computer-based assessment. *International Review of Ondustrial and Organizational Psychology*, 9, 31-69.

BPS (1995) *Psychological Testing: A User's Guide.* Leicester: British Psychological Society.

BPS (1999) *Guidelines for the Development and Use of Computer-Based Assessments*. Leicester: British Psychological Society.

Cisero, C.A., Royer, J.M., Marchant, H.G. and Jackson, S.J. (1997) Can the Computer-based Academic Assessment System (CAAS) be used to diagnose reading disability in college students? *Journal of Educational Psychology,* 89(4) 599-620.

College Board (1995) *Accuplacer 5.0 for Windows.* New York: College Entrance Examination Board.

Crook, C. (1994) *Computers and the collaborative experience of learning.* London: Routledge.

Crook, C. and Steele, J. (1987) Self-selection of simple computer activities by infant school pupils. *Educational Psychology*, 7, 23-32.

Dillon, A. (1992) Reading from paper versus screens: a critical review of the empirical literature. *Ergonomics,* 35(10), 1297-1326.

Dimock, P.H. and Cormier, P. (1991) The effects of format differences and computer experience on performance and anxiety on a computer-administered test. *Measurement and Evaluation in Counselling and Development,* 24, 119-126.

Dunn, L.M., Dunn, L.M., Whetton, C. and Burley, J. (1997) *British Picture Vocabulary Scale*. (Second Edition). Windsor, Berks: NFER-Nelson.

Durndell, A. (1991) The persistence of the gender gap in computing. *Computers and Education*, 16, 283-287.

Elliott, C.D., Smith, P. and McCulloch , K. (1996) *British Ability Scales (Second Edition).* Windsor, Berks: NFER-Nelson.

Evans, L.D., Tannehill, R. and Martin, S. (1995) Children's reading skills: a comparison of traditional and computerised assessment. . *Behaviour Research Methods, Instruments and Computers,* 27, 162-165.

French, C. (1986) Microcomputers and psychometric assessment. *British Journal of Guidance and Counselling,* 14(1) 33-45.

French, C. and Beaumont, J.G. (1990) A clinical study of the automated assessment of intelligence by the Mill Hill Vocabulary Test and the Standard Progressive Matrices test. *Journal of Clinical Psychology*, 46, 129-140.

Fuchs, L.S., Fuchs, D. and Hamlett, C. (1993) Technological advances linking the assessment of students' academic proficiency to instructional planning. *Journal of Special Education Technology,* 12, 49-62.

Fuchs, L.S., Fuchs, D. Hamlett, C. and Hasselbring, T.S. (1987) Using computers with curriculum-based monitoring: Effects on teacher efficiency and satisfaction.. *Journal of Special Education Technology,* 8, 14-27.

Godfrey Thompson Unit (1993) *Edinburgh Reading Tests.* London: Hodder and Stoughton.

Greenwood, C.R. and Rieth, H.R. (1994) Current dimensions of technology-based assessment in special education. *Exceptional Children*, 61(2) 105-113.

Hagues, N. and Courtenay, D. (1993) *NFER-NELSON Verbal Reasoning Test Series.* Windsor, Berks: NFER-Nelson.

Hambleton, R.K. and Swaminathan, H. (1985) *Item Response Theory: Principles and Applications.* Boston: Kluwer-Nijhoff.

Horne, J.K., Singleton, C.H. and Thomas K.V. (1999*) LASS Secondary Computerised Assessment System.* Beverley, East Yorkshire: Lucid Creative Limited.

Hughes, M. Brackenridge, A., amd Macleod, H. (1987) Children's ideas about computers. In J. Rutkowska and C. Crook (Eds.) *Computers, cognition and development.* Chichester: Wiley.

Maguire, K.B., Knobel, M.L.M., Knobel, B.L. and Sedlacek, L.G. (1991) Computer-adapted PPVT-R: a comparison between standard and modified versions within an elementary school population. *Psychology in the Schools,* 28, 199-205.

Martin, T.A. and Wilcox, K.L. (1989) Hypercard administration of a block-design task. *Behaviour Research Methods, Instruments and Computers,* 21, 312-315.

McCullough, C.S. and Wenck, L.S. (1984) Current microcomputer applications in school psychology. *School Psychology Review*, 13, 429-439.

Miller, L. DeJean, J. and Miller, R. (2000) The literacy learning curriculum and use of an Integrated Learning System. *Journal of Research in Reading*, 23(2), 123-135.

Moore, N.C., Summer, K.R. and Bloor, R.N. (1984) Do patients likes psychometric testing by computer? *Journal of Clinical Psychology,* 40, 875-877.

Naglieri, J.A. (1985) *Matrix Analogies Test.* New York: Psychological Corporation.

NCET (1994) *Using IT for Assessment – Going Forward*. Coventry: National Council for Educational Technology.

NCET (1995) *Using IT for Assessment – Case Studies Reports*. Coventry: National Council for Educational Technology.

NCET (1996) *Integrated Learning Systems – report of phase II of the pilot evaluation of ILS in the UK*. Coventry: National Council for Educational Technology.

Olsen, J.B. (1990) Applying computerized adaptive testing in schools. *Measurement and Evaluation in Counselling and Development*, 23, 31-38.

SCAA (1996) *Desirable Outcomes for Children's Learning on Entering Compulsory Education.* London: School Curriculum and Assessment Authority.

SCAA (1997) *The National Framework for Baseline Assessment: Criteria and procedures for Accreditation of Baseline Assessment Schemes.* London: School Curriculum and Assessment Authority.

Scott, T., Cole, M. and Engel, M. (1992) Computers and education: A cultural constructivist perspective. *Review of Research in Education*, 18, 191-251.

Singleton, C.H. (1997b) Computerised assessment of reading. In J. R. Beech and C. H. Singleton (Eds.) *The Psychological Assessment of Reading.* London: Routledge, pp. 257-278.

Singleton, C.H. and Horne, J.K. (in press) Computerised baseline assessment of mathematics. *Educational Research and Evaluation.*

Singleton, C.H., Horne, J.K and Thomas K.V. (1999) Computerised baseline assessment. *Journal of Research in Reading,* 22(1), 67-80.

Singleton, C.H., Horne, J.K. and Thomas, K.V. (in preparation) Prediction of literacy development from baseline measures on school entry.

Singleton, C.H., Horne, J.K. and Vincent, D. (1995*) Implementation of a Computer-based System for the Diagnostic Assessment of Reading*. Unpublished Project Report to the National Council for Educational Technology. Hull: Department of Psychology, University of Hull.

Singleton, C.H. and Thomas, K.V. (1994) Computerised screening for dyslexia. In C. H. Singleton (Ed.) *Computers and Dyslexia: Educational Applications of New Technology*. Hull: Dyslexia Computer Resource Centre, University of Hull, pp. 172-184.

Singleton, C.H., Thomas K.V. and Horne, J.K. (1998) *CoPS Baseline Assessment System.* Beverley, East Yorkshire: Lucid Research Limited.

Singleton, C.H., Thomas K.V., Leedale, R.C. and Horne, J.K. (in press*) Lucid Adult Dyslexia Screening (LADS).* Beverley, East Yorkshire: Lucid Creative Limited.

Singleton, C.H., Thomas K.V. and Horne, J.K. (2000) Computer-based cognitive assessment and the development of reading. *Journal of Research in Reading,* 23(2), 158-180.

Singleton, C.H., Thomas, K.V. and Leedale, R.C. (1996) *Lucid CoPS Cognitive Profiling System* Beverley, East Yorkshire: Lucid Research Limited.

Skinner, H.A. and Allen, B.A. (1983) Does the computer make a difference? Computerized versus face-to-face versus self-report assessment of alcohol, drug and tobacco use. *Journal of Consulting and Clinical Psychology,* 51, 267-275.

Smith, P. and Hagues, N. (1993) *NFER-NELSON Non-verbal Reasoning Test Series.* Windsor, Berks: NFER-Nelson.

Thomas K.V., Singleton, C.H, and Horne, J.K. (2001*) LASS Junior Computerised Assessment System.* Beverley, East Yorkshire: Lucid Creative Limited.

Thorndike, R.L., Hagen, E. and France, N. (1989) *Cognitive Abilities Test (Second Edition).* Windsor, Berks: NFER-Nelson.

Tymms, P. (1999) Baseline assessment, value added and the prediction of reading. *Journal of Research in Reading*, 22(1), 27-36.

Underwood, J. (2000) A comparison of two types of computer support for reading development. *Journal of Research in Reading,* 23(2), 136-148.

Underwood, J., Cavendish, S. Dowling, S., Fogelman, K. and Lawson, T. (1994*) Integrated Learning Systems in UK Schools.* Coventry: National Council for Educational Technology.

Underwood, J., Cavendish, S. Dowling, S., and Lawson, T. (1996*) Integrated Learning Systems: A study of sustainable gains in UK Schools.* Coventry: National Council for Educational Technology.

Wade, B. and Moore, M. (1993) The Test's the Thing: viewpoints of students with special educational needs. *Educational Studies*, 19(2), 181-191.

Watkins, M.W. and Kush, J.C. (1988) Assessment of academic skills of learning disabled students with classroom microcomputers. *School Psychology Review*, 17(1) 81-88.

Wesson, C., Fuchs, L.S., Tindal, G., Mirkin, P. and Deno, S. (1986) Facilitating the efficiency of on-going curriculum-based measurement. *Teacher Education and Special Education*, 9, 166-172.

Wilson, S.L., Thompson, J.A. and Wylie, G. (1982) Automated psychological testing for the severely physically handicapped. *International Journal of Man-Machine Studies*, 17, 291-296.

Wood, D., Underwood, J. and Avis, P. (1999) Integrated Learning Systems in the classroom. *Computers and Education,* 33(2/3), 91-108.

Woodward, J. and Rieth, H. (1997) A historical review of technology research in special education. *Review of Educational Research*, 67(4), 503-536.

Figure 1. Example item from the computerised Verbal Concepts test (“*Which picture goes best with the word ‘cooking’?”).*

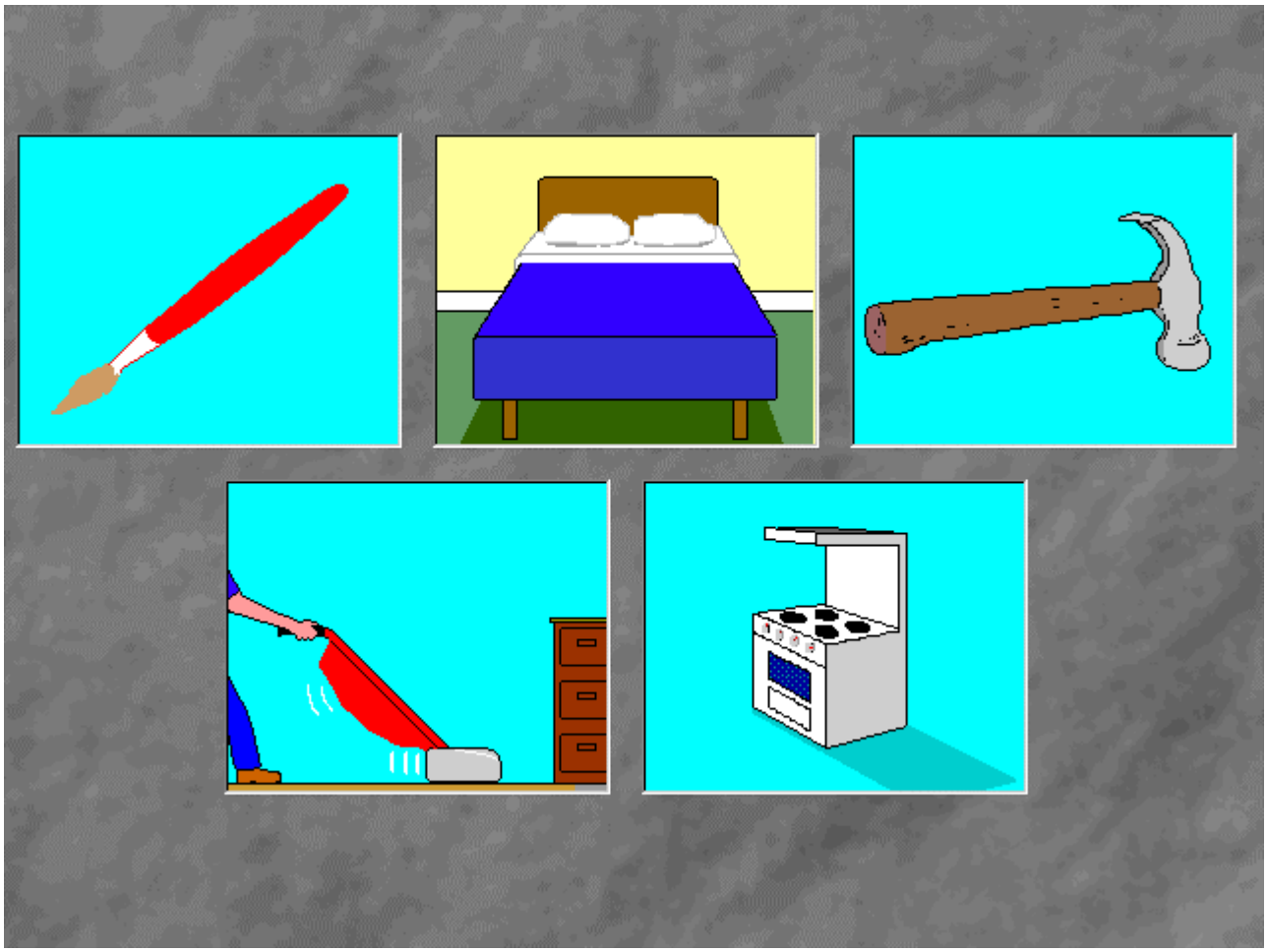

Figure 2. Example item from the computerised Mental Rotation test.

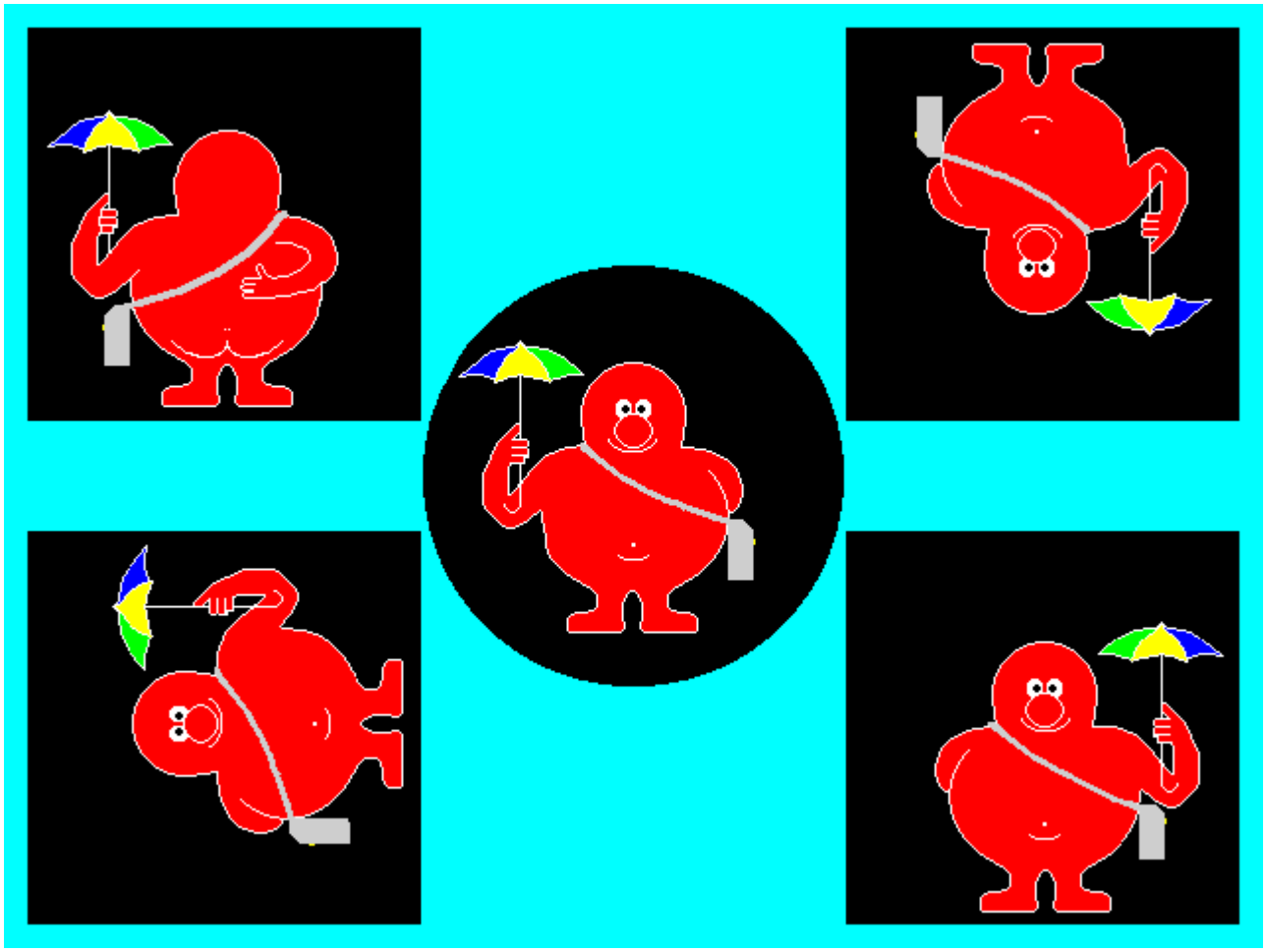

Table 1. Means and standard deviations of scores obtained in the computerised and conventional tests used in the study (N= 90 in each case).

Test	Score type	Mean	Standard Deviation
Verbal Concepts (CBA)	Raw score (out of 40)	25.86	3.30
Mental Rotation (CBA)	Raw score (out of 24)	9.16	4.35
Verbal Ability (BAS–II)	Standard score	107.02	11.40
Nonverbal Ability (BAS–II)	Standard score	106.97	16.34
Word Definitions (BAS–II)	T score	52.02	8.82
Verbal Similarities (BAS–II)	T score	55.96	6.85
Matrices (BAS–II)	T score	54.60	9.68
Quantitative Reasoning (BAS–II)	T score	53.39	10.71

Table 2. Correlations between the computer-based tests (raw scores) and BAS–II verbal ability and nonverbal ability (standard scores); N = 90.

	Verbal Concepts (CBA)	**Mental Rotation (CBA)**	**Verbal Ability (BAS–II)**
Mental Rotation (CBA)	0.33 $p < 0.002$		
Verbal Ability (BAS–II)	0.51 $p < 0.001$	0.13 (not significant)	
Nonverbal Ability (BAS–II)	0.30 $p < 0.004$	0.42 $p < 0.001$	0.40 $p < 0.001$

Table 3. Correlations between the computer-based tests (raw scores) and the BAS-II tests (ability scores); N = 90.

	Word Definitions	**Verbal Similarities**	**Matrices**	**Quantitative Reasoning**
Verbal Concepts	0.47 $p < 0.001$	0.52 $p < 0.001$	0.31 $p < 0.003$	0.26 $p < 0.02$
Mental Rotation	0.12 (not significant)	0.12 (not significant)	0.39 $p < 0.001$	0.38 $p < 0.002$

Table 4. Gender breakdown of results for the computer-based assessments (raw scores) and the BAS–II verbal and nonverbal ability measures (standard scores).

	Boys (N = 49)		**Girls (N = 41)**	
	Mean	Standard Deviation	Mean	Standard Deviation
Verbal Concepts (CBA)	25.78	3.60	25.92	3.07
Mental Rotation (CBA)	8.78	4.38	9.47	4.35
Verbal Ability (BAS–II)	105.98	11.15	107.90	11.65
Nonverbal Ability (BAS–II)	105.22	16.66	108.43	16.10

Table 5. Preferences expressed by boys and girls for the computer-based tests versus the conventional tests.

	Boys	**Girls**	**All**
Preferred computer tests	45	32	77
Preferred conventional tests	2	5	7
Don’t know / No preference	2	4	6
Totals	49	41	90