British Journal of Educational Technology *Vol 37 No 1 2006* 79–91
doi:10.1111/j.1467-8535.2005.00524.x

The equivalence of online and traditional testing for different subpopulations and item types

Robert MacCann

Dr Robert MacCann is head of Measurement and Research Services within the Board of Studies, in New South Wales, Australia. His research interests include the implementation of educational technology, item response theory, classical test theory and test equating. Address for correspondence: GPO Box 5300, Sydney, NSW, Australia 2001. Email: maccann@boardofstudies.nsw.edu.au

Abstract
A trial of pen-and-paper and online modes of a computing skills test was conducted for volunteer students of ages 15–16 in New South Wales, Australia. The tests comprised Matching, True/False and 4-option Multiple-Choice items. The aims were to determine whether gender, socioeconomic status (SES), or the type of item interacted with testing mode. No interactions were found for gender and item type, but the SES interaction was statistically significant. For low SES students, the online mode mean was 1 percent lower than the pen-and-paper mean, whereas high SES students had near equivalent means. These findings should be treated with caution as the groups in the study were self-selected, rather than random samples from the student population.

Introduction

In recent years educational measurement in Australia has been moving towards the use of computer-based testing. Computer-based tests are defined as tests or assessments that are administered by computer in either stand-alone or networked configuration, or by other technology devices linked to the Internet or the World Wide Web (Olsen, 2000). As testing shifts towards computer-based assessment, important practical issues arise that may affect the validity of the assessment. Currently in Australia, most major assessment systems use traditional formats in which examinees respond to printed examination papers by writing their answers in answer booklets (pen-and-paper testing). If examinees are given a choice whether to attempt the same examination by pen and paper or by computer-based testing, then equity requires that there be no advantage gained by attempting one mode over the other. In the *Standards for Educational and Psychological Testing*, Standard 6.11 states, 'If a test is designed so that more than one method can be used for administration or recording responses—such as marking responses in a test booklet, on a separate answer sheet, or on a computer keyboard—

then the manual should clearly document the extent to which scores arising from these methods are interchangeable' [American Educational Research Association (AERA), 1999, p. 70].

This paper is concerned with issues involved in assessing the equivalence of two modes of delivery of the Computing Skills Assessment (CSA) examination for Year 10 school students in New South Wales (NSW), Australia. The CSA examination arose from the paper *Plans for Education and Training, 1999–2003* (NSW Government, 1999) which foreshadows the introduction of state-wide, externally tested computing skills assessment for all Year 10 students. Initially, this testing would occur by means of an external pen-and-paper test. However, as the system evolves, the mode of testing would eventually shift entirely to that of computer-based testing. In the interim, the two modes of testing could coexist, with the pen-and-paper administration being the main mode of testing but with schools being given the option of having their students present for a computer-based test.

Background

There is a growing literature of research studies comparing the equivalence of computer-based and pen-and-paper tests. The following review will consider the main studies of relevance to the hypotheses in this paper. An early review, Mazzeo and Harvey (1988) compared the two modes of testing across a number of different testing domains. With some exceptions, the differences between the modes were generally small and not of practical significance. If one restricts the domain to that of achievement testing using short answer or multiple-choice items, which is the focus of the Computing Skills Assessment, then there seems to be little difference between the forms. Bunderson, Inouye and Olsen (1989) cite three studies of this type giving no significant differences. Mead and Drasgow (1993) reported a meta-analysis of 29 studies, finding only a small cross-mode effect size and concluding that there is no medium effect for 'carefully constructed power tests', while Bugbee (1996) concludes that testing by computer can be equivalent to a pen-and-paper test but that it is the responsibility of the test developer to establish that it is. This general conclusion that the differences across the two modes of testing are small should be restricted to short-answer and multiple-choice items. On extended-response items, for young people who have used computers throughout their schooling, testing via pen and paper yields underestimates of the students' skills in comparison to the same items requiring a response by computer (Russell, 1999; Russell & Haney, 1997; Russell & Haney, 2000).

While the above conclusion for short or multiple-choice items is generally based on studies involving overall populations (or samples from such populations), one may consider whether it holds equally well for certain subgroups. Important subgroups are the male and female subpopulations. Males and females have different patterns of interaction in the classroom with males likely to dominate computer use at school (Keogh, Barnes, Joiner & Littleton, 2000). An international comparative study of computer use found that females are less likely to have access to computers and that females exhibited higher levels of anxiety concerning computers and lower levels of confidence in oper-

ating them (Janssen Reinen & Plomp, 1993). Gallagher, Bridgeman and Cahalan (2002) found that for multiple-choice tests, females may receive higher scores on the pen-and-paper mode than on the computer-based mode. It is of interest to discover whether such differences in mean scores are also found in the Australian data and this is the first research question of interest: whether there is an interaction effect between sex and testing mode. That is, is the male/female mean score difference larger on one testing mode than the other?

A second set of subgroups which may perform differentially across the two testing modes are those based on socioeconomic status (SES). Schools may differ considerably in the type of computer education they can deliver and the level and sophistication of their hardware and software. These differences in computer access are also mirrored in the home, where computers may be entirely absent, or at the other extreme, used on an almost daily basis. In the US, Coley, Cradler and Engel (1997) found that the student:computer ratio is highest in schools with the largest proportions of poor and minority students and that internet access decreases as the percentage of such students increases. For internet access in US schools, a similar result was found by the National Center for Education Statistics (1999). For NSW schools, no centralised information is available on computer use.

Inequities in computer access may have both a direct and indirect impact on scores on computer-based tests relative to pen-and-paper tests. The direct effect simply reflects relative inexperience and unfamilarity in performing the operations required for the computer-based testing, which could introduce construct-irrelevant variance (Kirsch, Jamieson, Taylor & Eignor 1998; Taylor, Kirsch, Eignor & Jamieson, 1999). An indirect effect may manifest through affective responses such as computer anxiety, exacerbated by computer inexperience, which could influence computer-based test scores in a differential manner (see Marcoulides, 1988). On the other hand, Wise and Plake (1989) found nonsignificant effects for computer anxiety and computer experience on computer-based versus pen-and-paper testing modes. This motivates the second research question: is there an interaction between different SES groups and testing mode?

A third research area concerns the equivalence of performance between testing modes at an item level. The test comprised three item types: 4 matching items worth 5 marks each (a total of 20 marks), 14 true/false items worth 1 mark each (a total of 14 marks), and 66 multiple-choice items worth 1 mark each (a total of 66 marks). Apart from the studies by Russell and Haney (1997), which mainly focus on extended response essay items but did also consider multiple-choice and short-answer (but open-ended) items, there has been little research on how different types of objective items interact with the two testing modes. For example, the matching items were deliberately introduced into this trial as, in the online testing mode, they involve considerable manipulation of the 'mouse'. In this mode, the respondees initially read option A and then look at the picture of the application to see which of the numbered tags corresponds to this option. They must then move the mouse to the numbered tag, click on it and hold it down, and while holding it down, drag it across to fit into the option box. This operation has to be

performed five times, once for each option. An example of this type of item is shown in the Appendix.

The pen-and-paper mode, on the other hand, requires the student to shade in a bubble on a scannable paper answer sheet. The computer-based mode involves a considerable amount of mouse manipulation which then brings into play different degrees of expertise in this construct-irrelevant variable and differences in the functioning of the mouse. A student with debris in the mouse, causing it to stick, may take considerably longer to perform the necessary dragging. A third research question is then: is there an interaction between testing mode and item type? That is, do the mean score differences between item types differ across the testing modes?

The test and administration

The test was a purely objective test with a maximum possible total score of 100. It tested a number of areas in computing skills that students would have been exposed to in their general education up to Year 10. Section 1 comprised four matching items where for each item, the examinee had to match six numbered features on an image of an application on a monitor screen with five verbally descriptive options (A to E). Thus, one numbered feature would be left over. For each correct matching, 1 mark was awarded, making a maximum of 5 marks per item. Hence given the four items, Section 1 was worth 20 marks. Section 2 comprised seven parts with ten questions in each. The seven parts were: Word Processing, Spreadsheets, Databases, Multimedia, Graphics, Research and the Internet. The first two questions in each part were True/False (worth 1 mark each) and the remaining eight questions were 4-option multiple choice (worth 1 mark each). Therefore Section 2 was worth 70 marks. Section 3 comprised ten multiple-choice items (worth 1 mark each), giving a maximum possible mark of 10 for this section.

In the pen-and-paper (PP) mode, for each item, the examinees shaded the appropriate bubble on scannable answer sheets. For the online mode, examinees simply clicked the mouse on the appropriate answer box displayed on the monitor. For the matching items in the online version, one placed the cursor on the box for the selected answer and dragged it to its appropriate destination box (as described earlier). Navigation for the online test was simple—the questions were displayed at the bottom of the monitor screen in blocks of 20 items. To go to any question in the block, one only had to click on its box. To go to another block of 20 items, one simply clicked on an arrow pointing right to advance to later items, or on an arrow pointing left to go back to earlier items. Thus one could range over all items in the test and easily go back to check or correct earlier items. This navigation bar is shown at the bottom of the figure in the Appendix.

This trial of the two testing modes was conducted by the Board of Studies NSW, a state government body that conducts statewide testing for students in Year 10 (ages 15–16) and Year 12 (ages 17–18). Compulsory statewide tests in English, Mathematics, Science, History and Geography are held at the end of Year 10 and count towards the award of the School Certificate (SC). Whereas students in the above courses study a set

curriculum in each subject, there is no course in computing skills. Students are expected to acquire such skills from their general educational experiences over the range of courses they have taken during their education. The CSA program is experimental and participation was voluntary. Schools were invited to participate in the CSA program by responding to advertisement, yielding a total of 155 schools from a state population of 775 schools. Of these, 120 schools elected to do the test predominantly by pen and paper (PP), and 35 by the online mode. Both modes of testing were conducted on the same day but the online version was available for schools to download onto their school servers from a dedicated password-protected website, a week in advance of the testing date. Detailed instructions for installing the test and recommended procedures for conducting the test were available on this website. Improvements from the previous year in the online testing included the running of the test locally on the school server, with the facility to send the results back to the Board of Studies' computer system. If a school's Internet was congested or busy, the test could be taken offline.

The examinees

The examinees were from volunteer schools, and hence one might expect them to be possibly more able than the average Year 10 student in the NSW education system. As these volunteer students also had scores on the SC statewide test in Science, this could be investigated, as shown below in Table 1. This gives the summary statistics of each of the groups on the statewide test in Science, the SC test having the highest correlation with the CSA than other Year 10 tests.

These Science data show that the volunteer sample of 14 248 examinees (mean = 58.8) was slightly more able on SC Science than the population (mean = 56.4). This held for both males and females. In addition, it should be noted that the percentages of each sex volunteering for the CSA slightly differ from the population percentages, with relatively more females volunteering for the CSA (52.8% of group) despite comprising only 49.1% of the population.

Table 2 below shows the percentages of the sexes in the testing modes. This shows that females were even more strongly represented in the online testing group (55.6%) than in the PP group (52.5%), being over-represented in both groups compared to their percentage in the Year 10 population (49.1%).

Table 1: Summary statistics of Science marks for CSA volunteers and the population

	CSA volunteers			*Population*		
	N	*Mean*	*SD*	*N*	*Mean*	*SD*
Male	6722 (47.2%)	58.64	15.40	39882 (50.9%)	56.21	16.46
Female	7526 (52.8%)	58.99	15.09	38482 (49.1%)	56.50	15.86
Total	14248 (100%)	58.82	15.23	78364 (100%)	56.40	16.17

Table 2: Percentages of the sexes in the modes of testing

	Pen and paper	*Online*
Male	6079 (47.5%)	643 (44.4%)
Female	6722 (52.5%)	804 (55.6%)
Total	12801	1447

The above data indicates that the volunteer samples differ slightly in mean ability and in proportional representation from those that would have been obtained had random sampling been possible. This suggests that any findings based on these samples should be treated carefully.

Analysis

The need for a covariate

The online group was slightly more able than the PP group when measured on a covariate, SC Science. The average scores obtained by the PP and online testing groups on the CSA are shown on the left-hand side of Table 3 below. Focussing on the *left-hand side* and looking at the total group means, it appears that the online group (mean = 69.25) has performed slightly better than the PP group (mean = 68.65). It also appears that there could be an interaction effect between sex and the mode of testing in that the sexes are virtually equal on the PP test, but females have performed better on the online test with a mean of 70.50 compared to the male mean of 67.69.

Table 3: Mean scores on the CSA and Science tests

	CSA			*Science*		
	PP	*Online*	*Total*	*PP*	*Online*	*Total*
Male	68.66	67.69	68.56	58.64	58.63	58.64
Female	68.64	70.50	68.84	58.58	62.40	58.99
Total	68.65	69.25	68.71	58.61	60.73	58.82

However, a study of the *right-hand side* of Table 3 shows that these considerations are a function of the sample selection which necessitates a covariate be used for the analysis. First, the online group is better as a whole on Science, with a mean of 60.73 compared to the PP mean of 58.61. Secondly, the sexes attempting the PP test have very similar means on Science but in the online testing group, the females have outperformed the males on Science. Thus the apparent interaction pattern observed on the CSA is mirrored in the Science results and could be explained by the way the groups were self-selected.

Interaction of testing mode with sex and SES

To test whether there were any significant differences between the modes of testing and the performances of the sexes, an analysis of variance (ANOVA) using the General Linear Model with Science as a covariate was performed. In addition, a socioeconomic (SES) variable was introduced. This was derived from the home address postcode and was based on the general SES indicator provided by the Australian Bureau of Statistics. This measure was divided at the median into two categories, categorising a student as high SES or low SES. An alpha level for statistical significance was set at 0.05.

The first two research questions were investigated in this analysis. The first research question, concerning whether there was an interaction between sex and the mode of testing, gave a nonsignificant result ($p = 0.354$). That is, the sex differences in mean scores were not significantly different from one testing mode to the other.

The second research question was whether there was an interaction between SES status and mode of testing, that is, whether mean score differences between high SES students and low SES students were larger on one mode of testing than on the other mode of testing. The number of students in each category and the cell means are given below in Table 4. This interaction did prove to be statistically significant ($p = 0.032$) and so further investigation was conducted to estimate the size and direction of this effect.

Table 4: Mean scores for high and low SES groups on each testing mode

	Pen and paper		*Online*	
SES	*N*	*Mean*	*N*	*Mean*
Low	4893	67.3	482	66.5
High	5244	69.7	645	71.3

A regression analysis was run with the CSA score predicted by Science, testing mode and SES, with the significant SES by Mode interaction included. To aid the interpretation of this analysis, the variables had been coded as follows:

Mode:	0 = PP,	1 = Online
SES:	0 = Low,	1 = High.

The regression produced the following unstandardised coefficients shown in Table 5.

To gain an understanding of the impact of these regression coefficients, consider the four groups created by crossing the two levels of SES with the two modes of testing. These four groups are listed in the left-hand column of Table 6. Each of the other

Table 5: Unstandardised coefficients-dependent variable CSA

	Unstandardised regression coefficients	*Standard error*
(Constant)	38.208	.318
MODE	−1.089	.386
SES	.357	.162
SCIENCE	.512	.005
MODE*SES	1.302	.513

Table 6: Additive effects from each source in predicting a CSA score

Group	*Constant*	*Test Mode*	*SES*	*Mode by SES*	*Total*
Low SES, OL	38.2	−1.089	0	0	**37.1**
Low SES, PP	38.2	0	0	0	**38.2**
High SES, OL	38.2	−1.089	0.357	1.302	**38.8**
High SES, PP	38.2	0	0.357	0	**38.6**

columns gives the number of marks added to a regression prediction from that particular source, *as taken from Table 5*. For example, the constant of 38.2 is part of the regression line for all of the four groups. However, groups taking the online testing (OL) mode have a negative term (–1.089) added to their regression line predicted score, whereas the PP groups have zero. Similarly, high SES groups have 0.357 added to their regression line predicted score, but low SES groups have zero. Finally there is the interaction effect where only students in the high SES and online group have 1.302 marks added to their predicted score, while the other groups have zero. Summing across each row, the total impact from these sources is obtained.

From Table 6, there are four regression lines for predicting CSA from Science, one for each group as follows:

Low SES and Online
Predicted CSA = **37.1** + 0.512 Science.
Low SES and PP
Predicted CSA = **38.2** + 0.512 Science.
High SES and Online
Predicted CSA = **38.8** + 0.512 Science.
High SES and PP
Predicted CSA = **38.6** + 0.512 Science.

These regression lines differ only in the bolded constant term. They show that, for high SES groups, there is little difference in the predicted CSA mark, regardless of whether

the mode of testing is PP or Online. However, for low SES students, the PP examinees have a predicted CSA score about 1.1 marks higher than Online examinees ($p = 0.005$). For the Science covariate, an increase in one mark in Science provides a predicted increase in the CSA score of just over half a mark.

Interaction of testing mode with item type

In testing a possible interaction between item type and testing mode, three item types were investigated: Matching items, True/False items and Multiple-Choice items. As True/False items may be regarded as a subset of multiple-choice items, it seems unlikely that one would find that they behave differentially across testing modes. Hence one possibility would be to group them with the multiple-choice items and contrast this combined set with the matching items. However, the possibility of there being some unintended outcomes in the behaviour of True/False and Multiple-Choice items suggested that it was better to have three separate categories of item types. The Matching section was worth 20 marks, the True/False section was worth 14 marks and the Multiple-Choice section was worth 66 marks.

A repeated measures ANOVA was run to assess whether the different item types had a statistically significant interaction with the mode of testing. This proved nonsignificant ($p = 0.628$). The nonsignificance of this interaction can be seen clearly in Figure 1. In this graph, the mean score on the section for a particular item type is given on the vertical axis, while the horizontal axis shows the testing mode (PP = pen and paper, OL = online). The means are joined by straight lines to help the eye judge the degree of parallelism. For each pair being compared, the graphs exhibit the classic sign of

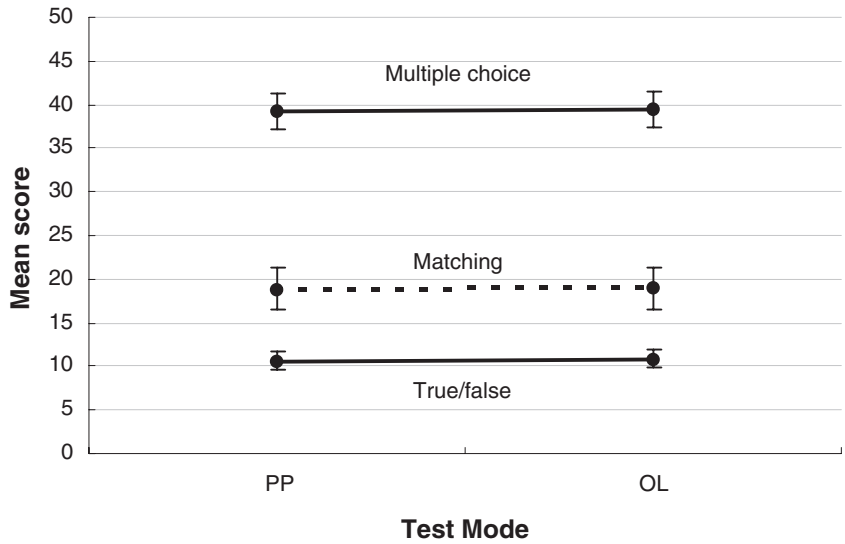

Figure 1: The noninteraction between testing mode and item type

noninteraction, the lines being nearly parallel, implying that the difference in the PP mode is nearly mirrored in the OL mode.

Conclusions

This paper investigated three research questions concerning the equivalence of the pen-and-paper and online modes of testing. The first question was concerned with whether males or females performed differentially across the two modes of testing. There is evidence that the sexes have different patterns of interaction in the classroom, with males attempting to get a greater share of computer use at school. It has also been found across a number of countries that females have shown higher anxiety levels concerning computer use and lower levels of confidence in operating them. One large study (Gallagher *et al,* 2002) found that females received higher scores in the PP mode rather than the computer-based mode for multiple-choice tests. However, for the NSW data involving testing on True/False, Multiple-Choice and Matching items, there was no significant interaction between sex and testing mode. One should, however, keep in mind the caveat expressed earlier that the male and female groups were self-selected, which may imply that the groups were somewhat atypical of those in the student population.

The second research question was whether the SES of a student interacted with the testing mode to produce higher average scores on one mode than the other. For the NSW data, a statistically significant interaction between testing mode and SES was found. In this case, an estimate of the effect showed that (all other things being equal) low SES groups would score about 1.1 marks (/100) higher on the PP mode than on the computer-based mode. High SES groups, on the other hand, had near identical predicted scores on the two modes. It is not implausible that low SES students, with less access to and experience with computers, would perform relatively better on the PP testing mode than on the computer-based mode. Factors that may be important here are the relative inexperience and unfamilarity in performing the operations required for the computer-based testing. In addition, affective responses, in part created by computer inexperience, could conceivably reduce scores differentially on the computer-based mode.

The argument above is that it is probably computer availability and its possible concomitant effects (such as computer unfamiliarity and increased test anxiety under testing by computer) that are responsible for the interaction. As data on student access to computers is not available, the variable SES is used as a proxy for student computer availability. It is possible, however, that it is not computer availability that accounts for the interaction between SES and testing mode. There may be another underlying variable responsible for this interaction of which we are unaware. The US data from Coley, Cradler and Engel (1997) and the National Center for Education Statistics (1999) implies that lower SES is associated with lower computer availability. This is also a plausible inference for the NSW system. If data on computer availability were at hand, a more direct analysis would be obtained by testing the interaction between computer availability (rather than SES) and testing mode. If computer availability were the key

factor in accounting for the significant interaction, then the small deficit for the low SES groups of 1 percent would be expected to diminish over time as personal computers become more freely available at home and in the school system.

The third research question was whether there is any interaction effect between the type of objective answer item and the mode of testing. The Computing Skills test comprised three-item types: Matching, True/False and 4-option Multiple Choice. The latter two categories would appear to be similar in the response movements required to answer the items in the computer-based mode, requiring minimal mouse movement. In contrast, the Matching items require considerable use of the mouse in the computer-based mode and, in theory, this could be a source of construct irrelevant variance. However, in practice these concerns were not borne out and the analysis showed no statistically significant interaction between item type and testing mode.

It is important to note the characteristics associated with this study that may have had an effect on the findings. First, the investigation is limited to the subject area of computing skills knowledge. This knowledge was not taught as part of a specific course, but was expected to be acquired by the students from their general educational experiences over a range of courses during their primary and high school education. Computing skills differ from the more traditional subject areas in that it is a rapidly changing one. Both these features suggest that the learning of such material may be more heavily influenced by experiences/resources outside the school than for the traditional subject areas. Second, the 155 schools taking part were volunteer schools from a population of 775 schools at the Year 10 level. On average, the volunteers were slightly more able than than the population and the disparity in the split between pen-and-paper and online volunteers could suggest that the latter were more experienced and confident in computing skills than a random sample would be. Third, slightly more females volunteered for the online test than would be expected from their representation in the population. It is possible that this group of females are more competent at computing skills than a random sample of such females, which could reduce the probability of finding gender differences. The use of a covariate (School Certificate Science) would be expected to ameliorate these effects, but this does not give the same degree of assurance that would have been obtained had random samples been possible. These factors suggest that the findings of this study should be treated with caution.

As computer-based testing becomes the dominant mode of testing, a practical problem may emerge in obtaining equity between forms. A phase will be entered where the pen-and-paper version becomes the 'back-up' form for schools, where only some form of misadventure has prevented students from taking the computer-based test. At the same time, the computer-based test would be attempting to move away from the traditional item types found in the tests in this paper. The new types of items in CSA would more fully imitate the real life actions that students would have to perform when using a computer. That is, the items would be more closely representative of an 'authentic assessment' of the tasks. When these types of questions are introduced, the ideal of retaining equivalence between a computer-based test and a pen-and-paper test will be

severely strained, as the emergence of authentic assessment may change the very construct that is being measured. Of course, when the pen-and-paper backup is no longer retained, this is no longer a problem. It is only a problem in the intermediate phase, where there is a tension between introducing new types of questions and retaining comparability between online and pen-and-paper tests.

References

American Educational Research Association (1999). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Bugbee, A. C. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education 28*, 282–299.

Bunderson, V., Inouye, D. K. & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement* 3rd ed (pp. 367–407). Phoenix, AZ: The American Council on Education and the Oryx Press.

Coley, R. J., Cradler, J. & Engel, K. (1997). Computers and classrooms: the status of technology in US schools. *Policy Information Report May 1997*. Princeton, NJ: Educational Testing Service.

Gallagher, A., Bridgeman, B. & Cahalan, C. (2002). The effect of computer-based tests on racial/ethnic and gender groups. *Journal of Educational Measurement 39*, 133–147.

Janssen Reinen, I. & Plomp, T. (1993). Some gender issues in educational computer use: Results of an international comparative survey. *Computers in Education 20*, 353–365.

Keogh, T., Barnes, P., Joiner, R. & Littleton, K. (2000). Gender, pair composition and computer versus paper presentation of an English language task. *Educational Psychology 20*, 33–43.

Kirsch, I., Jamieson, J., Taylor, C. & Eignor, D. (1998). Computer familiarity among TOEFL examinees. *TOEFL Research Report 59*. Princeton, NJ: Educational Testing Service.

Marcoulides, G. A. (1988). The relationship between computer anxiety and computer achievement. *Journal of Educational Computing Research 4*, 151–158.

Mazzeo, J. & Harvey, A. L. (1988). *The equivalence of scores from automated and conventional versions of educational and psychological tests: A review of the literature*. Research Report CBR 87-8, ETS RR 88–21. Princeton, NJ: Educational Testing Service.

Mead, A. D. & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis. *Psychological Bulletin 114*, 449–458.

National Center for Education Statistics (1999). Internet access. in public schools and classrooms: 1994-98. *Issue brief (NCES 1999-017)*. Washington, DC: NCES, US Department of Education.

NSW Government (1999). *Plans for Education and Training*, 1999–2003. Sydney, Australia, Department of Education and Training.

Olsen, J. B. (2000). *Guidelines for computer-based testing*. Retrieved May/June 2000, from http://www.isoc.org/oti/printversions/0500olsen.html.

Russell, M. (1999). Testing on computers: a follow-up study comparing performance on computer and on paper. *Education Policy Analysis Archives*. 7. Retrieved 14 February 2004, from http://epaa.asu.edu/epaa/v7n20/.

Russell, M. & Haney, W. (1997). Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper and pencil. *Educational Policy Analysis Archives,* 5. Retrieved 21 February 2004, from http://epaa.asu.edu/epaa/v5n3.html.

Russell, M. & Haney, W. (2000). Bridging the gap between testing and and technology in schools. *Educational Policy Analysis Archives, 8*. Retrieved 21 February 2004, from http://epaa.asu.edu/epaa/v8n19.html.

Taylor, C., Kirsch, I., Eignor, D. & Jamieson, J. (1999). Examining the relationship between computer familiarity and performance on computer-based language tasks. *Language Learning 49*, 219–274.

Wise, S. L. & Plake, B. S. (1989). Computer-based achievement. *Educational Measurement: Issues and Practice 8*, 5–10.

Appendix

Example of Online Screen Layout for a Matching Item