DIGITAL KOMPETANSE | 3–4-2009 | VOL. 4 | SIDE 143–158

Terje Manger | Ole-Johan Eikeland | Vibeke Vold
terje.manger@psysp.uib.no | efu99@hotmail.com | vibeke.vold@uib.no

A web-based national test of English reading as a foreign language: Does it test language ability, or computer competence?

A web-based measure of English reading and ICT competence

Abstract

The subjects were 389 nine to sixteen-year-old pupils who participated in the Norwegian web-based National Assessment in English Reading as a foreign language. The pupils' achievement in English, their self-reported Information and Communication Technologies (ICT) competence, and general ability were assessed. Using regression analyses, self-reported ICT competence, but not general ability scores, had a significant relation to English reading scores in the 4th grade but not in the 10th grade. In contrast, general ability, but not self-reported ICT competence, was significantly related to English reading for the 10th graders. The results at both grade levels were replicated in separate gender-wise analyses. The findings indicate that the computerized National Test in English reading in lower school grades is a test of computer competence or computer familiarity, more so than achievement in English. Thus, there are serious issues of validity involved when using computerized tests in primary schools.

KEYWORDS
National assessment • computerized tests • self-reported ICT competence • validity

Introduction

Does a computerized test in a school subject always measure what it is supposed to measure? Or can the test result be woven together with pupils' computer familiarity or computer competence? The purpose of the study was to investigate whether the web-based Norwegian National Assessment in English Reading really is a test of English reading, as opposed to simply reflecting students' computer competence. Computerized administra-

tion of achievement tests is increasing in frequency, thus it is particularly important to keep a critical attitude toward the confounding of the two kinds of skills when using such tests. In the present article, results from the web-based assessment will therefore be related to pupils' computer competence and to general ability, which is a criterion measure of achievement other than computerized test scores.

In 2005 the Norwegian government introduced standardized national tests and demanded obligatory national assessments for all pupils in reading and writing Norwegian, in mathematics, and in English reading for the 4th, 7th, and 10th grades in primary and lower secondary school, as well as the first year of upper secondary school. The tests in Norwegian, mathematics, and English writing were all paper-and-pencil-based while the English reading test was administered as a computerized test. In their evaluation report of the quality of test materials and results from the National Assessment in Norway, Lie, Hopfenbeck, Ibsen and Turmo (2005) argued that there is a high probability that pupils' experience with digital media can play a role in the validity of the computerized test. They claimed that, at least for the youngest pupils, there were large differences in relation to the handling of the PC, but the authors did not document their claim in the report.

The subsequent debate, in which many educators strongly opposed the existing national testing, together with a change of government at the end of 2005, led to the tests being put on hold. The critics focused on several problems, and pupils' familiarity with computerized tests was also commented on by critics other than Lie et al. (2005). In 2007, the National Assessments were improved and reintroduced by the new government. Again, the tests were conducted as a whole population approach, in mathematics and in reading (Norwegian and English) for grades 5 and 8, including computerized tests in English reading.

Traditional measures and computerized measures of achievement

In recent years a number of computerized educational tests have been introduced in several countries. However, the scientific literature shows that many problems exist in the use of computer-based testing of children, especially in the primary school. Obviously there is variability in both reading skills and computer experience. The web-based tests raise concerns that achievement in such subjects as mathematics or languages will be confounded with computer skills, introducing construct-irrelevant variance into ability measurement (Taylor, Kirsch, Eginor & Jamieson, 1999). That is, does the computer test measure what it is supposed to measure?

Several research reports have concluded that there are no significant differences between traditional achievement measures and computerized measures of achievement (Evans, Tannehill, & Martin, 1995; Taylor et al., 1999). Hargreaves, Shorrocks-Taylor, Swinnerton, and Threlfall (2004) found no relationship between performance on com-

puterized mathematics tests and competence with computers among ten-year-old children. The authors comment, however, that mathematics and computers have a strong link in schools, and that children should therefore already be familiar with working with computers in mathematics. They recommend that future studies focus on whether computer tests would be a fair manner of assessing children in other areas, such as in English. Contrary to the above results, Choi and Tinkler (2002) found that tests of reading were harder for third graders when presented on computers than when done via paper and pencil. Pomplum and Custer (2005) examined the score equivalence of computerized and paper-and-pencil scores for a series of reading screening tests, looking at nearly 2000 pupils at the levels of kindergarten through to the third grade. They found that computerized tests produced lower scores in reading than did a paper-and-pencil test. They suggest that school authorities should introduce computerized state assessments cautiously at the primary grades, especially for reading tests.

Scholastic achievement and ability

Can ability test scores, as assessed by verbal or non-verbal tests, be used as a criterion measure of achievement? The correlation between general intelligence and academic scores has been reported in the literature to be around r = .50 to .70 (Naglieri & Bornstein, 2003; Sattler, 2001). Further, non-verbal tests of general intelligence, such as the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test, have been found to be as predictive of achievement as those that also include verbal measures (Balboni, Naglieri & Cubelli, 2009). Naglieri (1996) administered a non-verbal test of general intelligence (Matrix Analogies Test-Short Form) to 2125 pupils in grades 2–9, as well as a measure of reading achievement (Multilevel Academic Survey Test-Reading). The average correlation between reading and intelligence was r = .57, with only minor variations due to age. The correlation between intelligence, as assessed by the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test, and mathematics and reading comprehension achievement test scores in Italian third- and fifth-year students varied between r = .32 and .52 (Balboni et al., 2009).

With respondents who have similar computer experience or competence, variability in such competence does not have an effect on scholastic achievement, while an accepted criterion measure of achievement, such as general ability, is expected to have an effect. If we assume that there are no significant differences between scores on traditional measures and computerized measures of achievement, intelligence may also be expected to be significantly related to scholastic achievement on computerized tests. Otherwise, if a computerized test and a paper-and-pencil test produce significantly different test results in such areas as foreign or native languages, then computer competence may be more strongly related to achievement on the computerized test than general ability. Likewise, if computer competence is more strongly related to test results than ability in such school

subjects, this may indicate that the computerized test is not a valid measure of achievement in those subjects.

Gender differences

The sex of the user is a primary individual difference variable of interest because of a number of findings regarding its influence on both ordinary school subjects and attitudes toward and interactions with computers. Gender differences in subjects favouring girls, particularly languages, are evident throughout the school years (e.g. Duckworth & Seligman, 2006). However, it was for a long period an established fact that there were gender differences in computer use, computer skills, and computer attitudes – favouring boys (Dundell & Thomson, 1997; Janssen & Plomp, 1997). Later studies (Kuhlemeier & Hemker, 2007; Volman, van Eck, Heemskerk & Kuiper, 2005) conclude, however, that gender differences in computer skills appear to be small. It is fair to assume that the gender differences in computer competence have shrunk over the years. The effect of utilizing computerized assessment on outcomes for the two sexes is thus not clear. Several studies have shown that females obtain lower scores on computerized assessments (Gallagher, Bridgeman, & Cahalan, 2002; Volman et al., 2005), while others have shown no differences (Horne, 2006). Horne found that girls outperform boys on paper-and-pen reading and spelling tests, but no significant gender differences were found on equivalent computerized tests. She suggests that it is likely that interaction with computers in the studies has enhanced the motivation of boys more than that of girls. This motivation effect could have compensated for, or masked, gender differences in performance. Other studies (Duckworth & Seligman, 2006) show that girls are more confident in answering questions about familiar material in school, but discouraged by unfamiliar problems presented in tests. Computerized tests may represent more unfamiliar material for girls than for boys, which also can mask gender differences in the subject measured.

Research questions

The following questions were raised for the present study: Will Information and Communications Technologies (ICT) competence have a different relation to computerized foreign language test scores in lower compared with higher school grades? Will there be gender differences? On the basis of previous research (Choi & Tinkler, 2002; Pomplum & Custer, 2005), and on the assumption that there would be less variability in ICT competence among pupils at a higher grade level, we had the following expectations: ICT competence will have a stronger relation to achievement on the computerized English reading test in the younger age groups, and general ability scores will have a stronger relation to English test scores in older age groups. We expected that there would be only minor gender differences.

Method

Subjects

During the spring of 2005 fifteen schools from both urban and rural areas were invited to participate in a study of age-related developmental aspects in technological interface design. Five schools from two rural municipalities (Fjell and Sund) and five schools from the city of Bergen accepted the invitation. The sample in the main study included 618 pupils: 191 pupils from the 4th grade (aged nine to ten), 187 pupils from the 7th grade (aged twelve to thirteen), and 240 pupils from the 10th grade (aged fifteen to sixteen). Of these pupils, 389 completed both the Norwegian National Assessment in English Reading, an Information and Communications Technology (ICT) competence questionnaire, and a non-verbal general ability test (117 pupils from the 4th grade, 124 from the 7th grade, and 148 from the 10th grade).

As a result of conflict about the use of national assessments between teachers, pupils, and parents, on the one side, and the Norwegian school authorities on the other side, some pupils boycotted the tests in 2005. At some schools only a portion of the pupils participated. This conflict had consequences for the internal response rate among the 618 pupils intended to take part in the study. In addition, at some schools the time schedule did not allow for both the national reading test and the ICT competence questionnaire to be performed on the same day, and some pupils did not show up on the following day to fill out the questionnaire. Moreover, some pupils filled out the questionnaire, but did not participate in the national reading test. There were also pupils who did not fulfil the test procedures to an acceptable degree; their test grades were therefore not given. In total 130 students occur as missing data in the English reading test. In addition, we demanded that the pupils should have completed at least ten of the seventeen ICT competence questionnaire items to be included in the index. Listwise deletion procedures among the variables in question account for the additional missing cases. Hence, the number of pupils included in the analyses was 389.

Instruments

The Norwegian National Test in English Reading is a test of English reading comprehension. It consists of a user interface that is quite similar for all age groups, but the difficulty levels for English reading vary across class levels. The test is linked to the competence goals in English, as outlined by the Norwegian Directorate for Education and Training. It measures pupils' abilities to find information (all grade levels), to understand the main content in texts (all grade levels), to reflect on the content in texts (7th and 10th grade levels), to choose different reading strategies depending on the purpose and situation (10th grade level), to read and understand texts of various lengths and genres (10th grade level), and to show the ability to distinguish between positively and negatively charged expressions referring to individuals and groups of people (10th grade level).

The pupils were to answer by clicking, dragging, and marking. The test formats were fixed to multiple-choices (clicking on check-off boxes, text boxes, and pictures), matching (dragging and dropping images in pictures), highlighting (marking a word), and colouring (clicking on paint colour and clicking in pictures). To develop the test system, the Norwegian Ministry of Education and the Norwegian Research Council founded the project group Bergen International Tests in English, Information Technology (BITE-IT), at the InterMedia department at the University of Bergen. The BITE-IT project was responsible for technical solutions and implementation of the computer-based test system.

Another project group, the Bergen International Tests in English (BITE), was responsible for the development of the English reading tasks to be used in the test. This group consisted mainly of English language experts. The BITE project group based its development of English reading tasks on the *Common European Framework of Reference for Languages* (Council of Europe, 2001). This framework defines levels of proficiency, allowing learners' progress to be measured at successive stages of learning on a life-long basis. It also provides the possibility of evaluating outcomes in an internationally comparable manner. Language competence was defined using three common reference levels: basic user, independent user, and proficient user. Hence, competence in English reading was graded at three levels, corresponding to the pupils' school levels. The top competence grade corresponds to a native speaker of that age.

Statisticians calibrated the test tasks for the various grades, and a designer produced the images to be used in the various test tasks. Two calibration tests, where all test tasks were tested with a representative sample of eleven thousand pupils, were conducted with the aim of testing the difficulty level of the tasks with regard to the CEF-level (Common European Framework) (Arntzen et al., 2004). An expert panel consisting of language experts and teachers thereafter selected the test tasks for the national tests, based on the calibration results and their own evaluations of the test questions. The alpha coefficients were α =.83, .75, and .84 for the 4th grade, 7th grade, and 10th grade, respectively (Lie et al., 2005).

A self-report ICT competence questionnaire carried out on paper was designed as part of the third author's dissertation research (Vold, 2007). It aimed to gather information about ICT competence, thought to be important in relation to pupils' mastery of interaction with the computer interface. A number of studies have revealed agreement between self-reports of ICT competence and other ICT competence scores (e.g. Sharma, 1991; Stefani, 1994). Self-reports may also provide insight into pupils' attitudes towards computing (Larres, Ballantine & Whittington, 2003). Self-reports have therefore been used extensively in the literature to assess computer knowledge and skills among pupils (Ballantine, Larres, & Oyelere, 2007). Although there are discussions about the reliability and consistency of self-reporting in the literature, self-reports are commonly accepted as a measure, and reliable and well-validated instruments are now frequently reported (Christensen & Knezek, 2008).

The self-report ICT competence measure was tested with pilot groups. The pupils were interviewed about how easily they understood these types of items. The final version contained 17 variables designed to measure how much pupils know within the following areas: use of the keyboard, mouse pointer, joystick, Internet (two variables), e-mail (two variables), creating own home page, talking with others on the computer (two variables), use of the computer to write letters, make graphics, use of diskettes, spreadsheets, search engines, find out who has created an image or writing something that he/she has found on the Internet, and download and install a program on the computer. Each item had five response categories (1 = lowest, 5 = highest competence). An index was created, containing the averaged sum of between ten (the least number of completed items acceptable) and 17 items related to pupils' self-reported ICT competence. Cronbach's alpha was $\alpha = .95$. The index covered the areas of digital literacy referred to as tool use, communication literacy, and web-literacy in the Digital Analysis Model as outlined by Ba, Tally, and Tsikalas (2002). All ICT concepts used were explained, often with examples (e.g. e-mail=use electronics mail to send letters on a computer). Both a first and a second version of the questionnaire were tested with several groups of pupils at the relevant ages.

The Matrix Analogies Test, Short Form (MAT-SF), is a general ability test that measures non-verbal reasoning. While Raven's Progressive Matrices, which have also been used for this purpose, lack U.S. norms, the Matrix Analogies Test was developed to meet the need for a well-normed and well-constructed test of non-verbal reasoning ability. The MAT-SF can be administered as a screening test in a group setting (Naglieri, 1985a), and uses 35 abstract designs of the standard progressive matrix type (a maximum score is 35 and a minimum score is 0). The standard progressive matrix format provides a measure of general reasoning ability and is especially useful when a person has limited language skills. The MAT-SF is highly related to non-verbal ability as measured by the Wechsler Intelligence Scale for Children, WISC-R ($r = .68$, $p<.001$) (Naglieri, 1985a). According to Prewett, Bardos, and Naglieri (1988), the test correlates significantly with all areas of achievement in school subjects. There is a 25-minute time limit for the test. Naglieri (1985b) reports an internal consistency median coefficient (alpha) across grades of $\alpha = .83$. In our study the alpha coefficients were $\alpha = .90$, .81, .84, and .87 for the 4th grade, 7th grade, 10th grade, and across grades, respectively.

Procedure

The testing period for the National Assessment in English Reading in 2005 was scheduled for three weeks during January and February for the 7th grade, three weeks in February and March for the 10th grade, and three weeks in April and May for the 4th grade. The English reading test, which was computer-based, was administered by the teachers. The self-report ICT competence questionnaire and the MAT-SF were administered in the classes by the third author within the same month as the English test. The survey was

approved by the Norwegian Data Inspectorate in accordance with Norwegian law. The Inspectorate asked the research group to obtain active consent from the parents of the pupils in school years 4 and 7, but not from parents of pupils in year 10, who were only informed about the purpose and procedure.

Analyses

Means, standard deviations, correlations, and t-tests were calculated. The ordinary least square method (OLS) was used with simultaneous inclusion of variables. Regression analyses were done separately for all three grades, in addition to gender-wise analyses within grades. SPSS, 15.0.1, was used for all analyses.

Results

Table 1 shows means and standard deviations gender-wise on the Information and Communications Technology (ICT) competence index, the MAT-SF, and the National Test in English Reading. The ICT competence index is the arithmetic mean of the 17 items. Given the answering format, this produces an average score between 1 (lowest competence) and 5 (highest competence).

Table 1. Girls' and boys' means, t-values for differences between boys and girls, and standard deviations on the Information and Communications Technology (ICT) competence index, Matrix Analogies Test, Short Form (MAT-SF) scores, and National Test in English Reading scores.

	Mean		Standard deviation			N	
	Girls	Boys	Girls	Boys	T-value	Girls	Boys
4th grade							
ICT competence	2.2	2.4	0.90	0.93	1.10	57	60
MAT-SF	20.7	20.8	6.45	6.88	0.84		
English Reading	2.4	2.3	0.88	0.90	0.61		
7th grade							
ICT competence	3.0	2.9	0.85	0.93	1.32	72	52
MAT-SF	27.4	26.7	4.47	5.08	0.89		
English Reading	5.2	5.1	0.81	0.75	1.26		
10th grade							
ICT competence	3.4	3.7	0.78	0.82	1.93	72	76
MAT-SF	30.5	28.8	3.35	5.08	2.45*		
English Reading	6.9	6.7	1.42	1.53	1.11		

* $p < .02$ (two-tailed).

A t-test showed that 10th grade girls had significantly higher MAT-SF scores than boys in the same grade (t=2.45, p <.02). No other significant differences between boys' and girls' mean scores within grades were evident.

Table 2 presents correlations between the ICT competence index, MAT-SF scores, and English Reading scores.

Table 2. Bivariate correlations between the ICT competence index, Matrix Analogies Test (MAT-SF) scores, and National Test in English Reading scores.

	ICT competence			MAT-SF		
	Boys	Girls	All	Boys	Girls	All
4th grade						
MAT-SF	.11	.24*	.17*	–	–	–
English Reading	.38**	.41**	.39**	.25*	.18*	.22**
7th grade						
MAT-SF	.25*	–.27**	–.02	–	–	–
English Reading	.36**	.05	.19*	.21	.17	.19*
10th grade						
MAT-SF	.08	–.34**	–.11	–	–	–
English Reading	–.06	–.04	–.06	.44**	.35**	.41**

**p < .01; *p < .05.

In the 4th grade there were significant positive bivariate correlations between scores on the English reading test and both self-reported ICT competence and MAT-SF scores (Table 2). This held for both sexes. However, in the 7th grade, only the boys' scores on English reading and ICT competence correlated significantly (r=.36). At the 10th grade there were no significant correlations between English reading and self-reported ICT competence, no matter whether for all pupils or gender-wise. The correlation between MAT-SF and English reading was, however, stronger for the 10th grade (r=.41) than for the 4th grade (r=.22), but Fisher's Z-test (Knoke & Bohrnstedt, 1994) shows that the difference between the two correlations is not significant (two-tailed, Z=1.88, n.s.). (A one-tailed null hypothesis that the correlations are equal can be rejected, p < .05.) According to Cohen's (1969) guidelines, the correlation for the 4th grade is small, while the correlation for 10th grade is moderate. The correlation between girls' scores on MAT-SF and ICT ranged from small positive (r=.24) in the 4th grade to small negative (r=• .27) in the 7th grade. The correlation remained negative (r=• .34) in 10th grade.

Table 3 presents the results from regression analyses, where the English reading test score is the predicted variable in the analyses.

Table 3. Results of regressing ICT competence index and Matrix Analogies Test (MAT-SF) scores on pupils' National Test in English Reading scores. Separate analyses for grades, and gender within grades. OLS done simultaneously. Two-tailed significance level.

		B	Se B	®
4th grade				
All pupils	ICT competence	.35	.08	.36**
	MAT-SF	.02	.01	.16
Girls	ICT competence	.38	.13	.39**
	MAT-SF	.01	.02	.09
Boys	ICT competence	.35	.12	.36**
	MAT-SF	.03	.02	.21
7th grade				
All pupils	ICT competence	.17	.08	.19*
	MAT-SF	.03	.02	.20*
Girls	ICT competence	.10	.12	.11
	MAT-SF	.04	.02	.20
Boys	ICT competence	.26	.11	.32*
	MAT-SF	.02	.02	.13
10th grade				
All pupils	ICT competence	–.03	.14	–.02
	MAT-SF	.14	.03	.41**
Girls	ICT competence	.17	.22	.09
	MAT-SF	.16	.05	.38**
Boys	ICT competence	–.18	.20	–.09
	MAT-SF	.14	.03	.45**

Note: In 4th grade adj. R^2 = .162, .147 and .162 for all pupils, girls and boys, respectively. In 7th grade adj. R^2 = .058, .012 and .108 for all pupils, girls and boys, respectively. In 10th grade adj. R^2 = .155, .102 and .182 for all pupils, girls and boys, respectively.
* p <.05; ** p < .01.

Self-reported ICT competence, but not MAT-SF scores, had a significant relation to English reading scores in 4th grade. However, the analyses of the 10th grade pupils showed quite opposite results: self-reported ICT competence had no significant relation to English reading, but the scores on MAT-SF were related significantly to English reading. In

the 7th grade both self-reported ICT competence and MAT-SF scores were related significantly to English reading. Neither self-reported ICT competence nor MAT-SF scores were, however, related to girls' English reading scores at this level, when gender-wise analyses were done. In contrast, there was a significant relation between boys' self-reported ICT competence and their English scores. All regression analyses (done separately for grades, and gender within grades) showed significant F-values.

Discussion and conclusions

In the present study achievement in the Norwegian web-based National Assessment in English Reading as a foreign language, self-reported ICT competence, and general ability were assessed. Since agreement between self-evaluation and alternative measures of ICT competence has been reported in a number of studies (e.g. Ballantine et al., 2007; Christensen & Knezek, 2008; Sharma, 1991; Stefani, 1994), a self-report measure was regarded as acceptable in the current study. Likewise, due to the high correlation between general ability and achievement (e.g. Naglieri & Bornstein, 2003; Sattler, 2001), an ability test was used as a criterion measure of achievement.

Self-reported ICT competence, but not general ability scores, had a significant relation to English reading scores in the 4th grade, but not in the 10th grade. In contrast, general ability, but not self-reported ICT competence, was significantly related to English reading for the 10th graders. The findings are in line with our hypotheses and consistent with those of other studies that question the validity of computerized achievement tests at lower primary school grades (Choi & Tinkler, 2002; Pomplun & Cluster, 2005). The computerized Norwegian National Test in English reading, examined in this study, appears to be well-confounded with computer skills, particularly for the younger pupils. We hypothesized that there would be only minor gender differences. However, in the 7th grade, the significant correlation between English reading and self-reported ICT competence remained for boys, but not for girls.

The Norwegian school authorities decided that the National Assessment in English Reading as a foreign language was available only per computer, and a paper-and-pencil test was not constructed specially for the study. The computer-based test was designed as an adaptive test, whereby the difficulty level of a given test question depended on the pupil's previous answers. The fact that the computer-based tests were adaptive made it quite difficult to construct a paper-based version of this test that could be used for research purposes. Thus, it was not feasible to conduct an experimental design with two groups (one using paper and the other using computer). The Matrix Analogies Test (MAT-SF) (Naglieri, 1985a), a general ability test that measures non-verbal reasoning, was therefore used as a criterion measure of achievement.

Intelligence has a significant relation to scholastic achievement (e.g. Prewett et al., 1988; Sattler, 2001), and in several studies general ability, also assessed by non-verbal tests, has been used as a criterion measure of achievement in school subjects (e.g. Naglieri, 1996) or to predict academic achievement (e.g. Gustafsson & Undheim, 1996). This does not mean that these tests should replace achievement tests for measuring present achievement, but that they may provide information that is useful for predicting achievement, also in language cultures other than the English (Balboni et al. 2009). Non-verbal tests of general ability are more appropriate for assessing children with different levels of language, knowledge, and cultural backgrounds. Some correlations between achievement in school subjects and non-verbal ability in non-English cultures are, however, smaller, but the magnitude of these values in contrast to previous findings may be related to differences in the achievement tests studied (Balboni et al., 2009). Although there is a lack of studies looking at correlations between non-verbal tests of general ability, such as the MAT-SF, and achievement in foreign languages, it is concluded that the non-verbal ability tests appear to be useful in educational settings across cultures (Naglieri, 1985b; Naglieri & Ronning, 2000).

Familiarity with using computers can well lead to increments in performance on computer-assisted or computer adaptive tests (Kirsch, Jamieson, Taylor & Eignor, 1998; McCullough, 1990). Thus, in discussing validity, where the key issue is if one is able to measure what one wants to measure (construct validity), it is particularly important to be aware of the possible confounding of two kinds of skills when using computer-based assessment. Especially for the younger pupils, there are large differences in relation to the handling of the PC, which was also commented on by Lie et al. (2005) in their evaluation report of the National Assessment in Norway. If these differences are not compensated for by training, the most computer-experienced pupils will outperform others on computerized tests of achievement.

Were pupils trained specially for the Norwegian computerized National Assessment or otherwise prepared for the tests? To narrow the gap, the project group that designed the computer-based test prepared demonstration test tasks for all of the grade levels that were being tested. They also sent information letters to the schools recommending that pupils practise with the demonstration test tasks. It was argued that the pupils would benefit from familiarizing themselves with the interface modalities and the format of the test tasks. However, responses from schools, when the national assessment was introduced in Norway, indicated that many of the teachers had little time to carry out the intended national test preparation activities, those which could potentially reduce the differences. Teachers commented that the national tests were introduced after the planning of the school activities was finished, making it difficult to find time for including such test activities.

The Norwegian National Assessment in English Reading intended to measure the pupils' basic abilities in that language. The data from the introductory year of national assessments in Norway show clearly that the test results in lower primary schools seem to be woven together with pupils' computer skills. What is then being measured? The pupils'

proficiency in English, or facility in using computer technology? The pupils' scores in English Reading in lower primary school may well be a test of computer competence or computer familiarity, much of this acquired inside or outside school, more than achievement in English. In other words, the test does not necessarily measure what it is supposed to measure, as observed earlier by Lie et al. (2005).

A curious finding is the significant negative correlation between girls' MAT-SF scores and their self-reported ICT competence in the 7th and 10th grades. The finding is difficult to explain, but may suggest that the girls with high MAT-SF scores in this cohort, within their school years, developed negative attitudes toward computers and regarded it as "boyish" and less relevant for school subjects. Further studies are needed to uncover whether the finding is confirmed in future cohorts of girls. A possible future reduction of gender-wise differences in computer attitudes and competence may influence the correlation between general ability and self-reported ICT competence.

In conclusion, the study indicates that lack of computer competence or computer familiarity has an effect on computerized test results on achievement tests, particularly with the English language test. If computers are to be used for providing fair assessments of achievement, then care must be taken to ensure that all pupils have sufficient experience with computers or that training is provided in school. On the other hand, it seems likely that also very young pupils' access to and experience with computers will continue to increase, in home and school, and thus, differences between the pupils will possibly diminish over time. Although the studies by Choi and Tinkler (2002), Pomplun and Custer (2005), and our own study indicate that computerized tests of school achievement do not necessarily measure the content they are supposed to measure, further research is needed. Researchers should continue to investigate the factors behind the primary grade score differences, especially computer familiarity and also differences between test administration by teachers and computerized administration. Further studies should also compare paper-and-pencil and computer reading score differences, both in English as a foreign language, native languages, and in other subjects. Especially, if web-based compulsory nationwide tests in subjects such as native and foreign languages are to become more prevalent, the issues of construct validity need to be taken seriously. If such tests are to be a fair way to assess achievement, then care must be taken to ensure that the tests are measuring both representative aspects and the depths of the school subjects.

Acknowledgements

The third author's work was funded by a grant to Professor Konrad Morgan from the Norwegian Research Council and supported by the Faculty of Social Sciences at the University of Bergen. The authors would like to thank Professor Robert A.Wicklund for reading and commenting on an earlier draft of this manuscript.

References

Arntzen, M. E., Dragsnes, S., Hansen, C., Hansen, O., Morgan, K., von Schlanbusch, H. & Schlanbusch, L. (2004). *Erfaringsrapportert IKT-baserte nasjonale prøver i engelsk* [Experience reported ICT-based national tests in English]. BITE[IT] Bergen interaktive tester i engelsk.

Ba, H., Tally, W., & Tsikalas, K. (2002). Investigating children's emerging digital literacies. *The Journal of Technology, Learning and Assessment*, *1*, 1–49. Retrieved, June 5, 2007, from http:/www.bc.edu/research/intasc/jtla/journal/pdf/v1n4_jtla.pdf

Ballantine, J.A, Larres, P. M., & Oyelere, P. (2007). Computer usage and the validity of self-assessed computer competence among first-year business students, *Computers & Education*, *49*, 979–990.

Balboni, G., Naglieri, J. A., & Cubelli, R. (2009). Concurrent and predictive validity of the Raven Progressive Matrices and the Naglieri Nonverbal Ability Test. *Journal of Psychoeducational Assessment. OnlineFirst.* First published on August 25, 2009 as doi:10.1177/0734282909343763.

Choi, S. W., & Tinkler, T. (2002, April). *Evaluating comparability of paper-and-pencil and computer-based assessment in a K-12 setting*. Paper presented at the annual meeting of AERA, New Orleans.

Christensen, R., & Knezek, G. (2008). Self-report measures and findings for information technology attitudes and competencies. In: J. Voogt & G. Knezek (eds.): *International handbook of information technology in primary and secondary education* (pp. 349–365). New York: Springer.

Cohen, J. (1969). *Statistical power analysis of the behavioral sciences*. New York & London: Academic Press.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Retrieved June 5, 2007, from *http://www.coe.int/T/DG4/Linguistic/Source/Framework_EN.pdf*

Duckworth, A. L., & Seligman, M. E. P. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades and achievement test scores. *Journal of Educational Psychology, 98, 198–208*.

Evans, L. D., Tannehill, R., & Martin, S. (1995). Children's reading skills: A comparison of traditional and computerized assessment. *Behavior Research Methods, Instruments, and Computers, 27*, 162–165.

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.

Gustafsson, J.-E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In: D. C. Berliner & R. C. Calfee (eds.): *Handbook of educational psychology* (pp. 186–242). New York: Prentice Hall.

Hargreaves, M., Shorrocks-Taylor, D., Swinnerton, B., Tait, K., & Threlfall, J. (2004). Computer or paper? That is the question: does the medium in which

assessment questions are presented affect children's performance in mathematics? *Educational Research, 46,* 29–42.

Horne, J. (2006). Gender differences in computerized and conventional educational tests. *Journal of Computer Assisted Learning*, *23*, 47–55.

Janssen, R. I., & Plomp, T. J. (1997). Information technology and gender equality: a contraICTtion in terminis. *Computers and Education*, *28*, 65–78.

Kirsch, I., Jamieson, J., Taylor, C., & Eignor, D. (1998). *Computer familiarity among TOEFL Examinees. TOEFL Research Report*. 59. Princeton, N.J.: Educational Testing Service. Retrieved January 6, 2007, from *http://ftp.ets.org/pub/toefl/275755.pdt*

Knoke, D. & Bohrnstedt, G. W. (1994). *Statistics for social data analysis*. Third ed. Itasca, Illinois: F. E. Peacock Publishers, Inc.

Kuhlemeier, H. & Hemker, B. (2007). The impact of computer use at home on students' internet skills. *Computers and Education*, *49*, 460–480.

Larres, P.M., Ballantine, J.A., & Whittington, M., (2003). Evaluating the validity of self-assessment: measuring computer literacy among entry-level undergraduates within accounting degree programmes at two UK universities, *Accounting Education, 12*, 97–112.

Lie, S., Hopfenbeck, T. N., Ibsen, E., & Turmo, A. (2005). *Nasjonale prøver på prøve. Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater på nasjonale prøver våren 2005* [National Tests on test. A report from a committee that has evaluated the quality of tasks and results of the National Test in the Spring of 2005]. Oslo: Institutt for lærerutdanning, Universitetet i Oslo.

McCullough, S. (1990). Computerized assessment. In: Reynolds, C. R. and Kamphaus, R. W. (eds.): *Handbook of Psychological and Educational Assessment of children, intelligence and achievement* (pp. 723–747). New York: Guilford Press.

Naglieri, J. A. (1985a). *Matrix Analogies Test. Short Form*. New York: The Psychological Corporation/Harcourt Brace Jovanovich, Inc.

Naglieri, J. A. (1985b). *Matrix Analogies Test. Short Form. Examiner's Manual*. Columbus: Charles E. Merill Publishing Company and A. Bell and Howell Company.

Naglieri, J. A. (1996). An examination of the relationship between intelligence and reading achievement using the MAT-SF and MAST. *Journal of Psychoeducational Assessment*, *14*, 65–69.

Naglieri, J. A., & Bornstein, B. T. (2003). Intelligence and achievement: Just how correlated are they? *Journal of Psychoeducational Assessment*, *21*, 244–260.

Naglieri, J. A., & Ronning, M. E. (2000). Comparison of White, African-American, Hispanic and Asian children on the Naglieri Nonverbal Ability Test. *Psychological Assessment, 12,* 328–334.

Pomplun, M., & Cluster, M. (2005). The score comparability of computerized and paper-and-pencil formats for K-3

reading tests. *Journal of Educational Computing Research, 32,* 153–166.

Prewett, P. N., Bardos, A. N., & Naglieri, J. A. (1988). Use of the Matrix Analogies Test – Short Form and the Draw-a-Person: A quantitative scoring system with learning-disabled and normal students. *Journal of Psychoeducational Assessment, 6,* 347–353.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego, CA: Jerome M. Sattler, Inc.

Sharma, N. K. (1991). How accurate are human beings in their self-assessment? *Psychological Studies, 26*, 77–82.

Stefani, L. A. J. (1994). Peer, self and tutor assessment: relative reliabilities. *Studies in Higher Education, 19*, 69–75.

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.

Vold, V. (2007). Age-related development aspects in technological interface design. Dissertation for the degree of philosophiae doctor (PhD). Unpublished. University of Bergen, Norway.

Volman, M., van Eck, E. Heemskerk, I., & Kuiper, E. (2005). New technologies, new differences. Gender and ethnic differences in pupils' use of ICT in primary and secondary education. *Computers and Education, 45*, 35–55.