Text on Paper; Text on Screen. Does it Matter?

Keith Lenden-Hitchcock and Syed Mohamed
Singapore Examinations and Assessment Board

Abstract

The assessment of oral communication where certain language skills such as reading and speaking may be assessed, are usually done through components such as Reading Aloud, Picture Discussion and Conversation. Reading skills tested in the Reading Aloud component are usually assessed by having candidates read aloud a text printed on paper. Does it matter if candidates read a text off computer screen or off printed paper or will it make a difference to candidates' performance? A research study was carried out on a sample of Grade 10 students (15 to 16 years of age) to investigate the hypothesis that "The computer, used as a replacement for text printed on paper, has no impact on the performance of students in an Oral Communication Examination. This paper describes the design and conduct of the research study and presents its findings. It will also discuss the implications of the `findings for high stake and large scale examinations.

Focus of Paper

This paper presents the findings of a preliminary study on conducting a component of the oral examination using computer. Grade 10 students from three courses of study involving a total of 636 students from 7 secondary schools in Singapore sat two separate oral examinations - one where they were asked to read aloud a passage printed on paper and the other to read aloud another passage displayed on a computer screen.

Background

In Singapore, every student has the opportunity to undergo at least ten years of general education. Almost all children start their formal education from age 6, spending six years in primary schools. At the end of Grade 6 they will sit their first major national examination before proceeding to do a 4 or 5-year secondary education at the end of which they will sit the Singapore-Cambridge General Certificate of Education (GCE) Normal (Technical), Normal (Academic) or Ordinary Level examinations.

The Oral Communication Examination

The GCE English Language examinations typically comprise three to four papers testing writing skills, language use and comprehension skills, listening skills and oral communication skills.

Oral Communication

A typical Oral Communication paper will consist of two or three sections, namely, Reading Aloud, Picture Discussion and Conversation. For Reading Aloud students are required to read a passage of about 250 words and they are assessed on three sets of criteria – (1) pronunciation and articulation, (2) rhythm and fluency and (3) expressiveness. For Conversation, students talk on a topic given by the examiner and they are assessed on their ability to give personal responses to the topic; their ability to speak clearly and confidently using appropriate vocabulary and accurate language; and their ability to interact with the examiner during the conversation.

The Study

The study is focussed on one of the components of the Oral Communication examination – Reading Aloud. The purpose of this study is to test the null hypothesis that "the computer has no effect on

students' performance", or "the mode of delivery (paper or computer) has no impact on student performance in the Reading Aloud component of the oral examination." Results from each course were analysed separately – only results from the Normal (Academic) and Express Courses are discussed here, though observations on those two courses apply equally to the Normal (Academic) course pupils.

Methodology

Two Equal Groups

Within each school, students from the same course were divided into 2 equal groups, Group I and Group II. These groups were considered equal in that within-group distributions of the following combinations of background variables were the same[1]:

- Gender
- Ethnic group
- Socio-economic status[2]

In addition, the number of students in each group was the same.

Four-Stage Examination

Each student from either group undertook a four-stage examination on two texts, Text 1 and Text 2.

i) 5 minutes reading preparation of Text 1
ii) 4 minutes examination on Text 1
iii) 5 minutes reading preparation of Text 2
iv) 4 minutes examination on Text 2

While the order of texts – Text 1 followed by Text 2 – was the same for either group, the delivery mode of the text, computer or paper, was different for the two groups in the following way: Group I students completed Stages (i) and (ii) on computer, followed by Stages (iii) and (iv) on paper. Group II students, on the other hand, first completed Stages (i) and (ii) on paper, followed by Stages (iii) and (iv) on computer.

Further, the team of examiners who examined students on Text 1 was not the same team as those who examined them on Text 2. These two examiners teams are labelled Team 1 and Team 2.

The above discussion can be summarised in the following way:

Group	Text 1/Team 1	Text 2/Team 2
I	(i) 5 minutes preparation on computer. (ii) 4 minutes examination on computer.	(iii) 5 minutes preparation on paper. (iv) 4 minutes examination on paper.
II	(i) 5 minutes preparation on paper. (ii) 4 minutes examination on paper.	(iii) 5 minutes preparation on computer. (iv) 4 minutes examination on computer.

[1] For the Normal (Technical) Course only Gender and Ethnic group were used to define equality.
[2] The measure of socio-economic status (SES) used here is a function of an individual's parents' education and housing type.

Considerations and Assumptions

On Texts

For each group of students, two appropriate texts, Texts 1 and 2 that are of comparable difficulty were selected from past GCE Ordinary Level papers for the students from the Express course. Similarly two comparable texts were selected from the GCE Normal (Academic) paper for the students from the Normal (Academic) course.

On Examiners

There are two different teams of oral examiners – Teams 1 and 2. While they may have different experiences, they were taken through standardisation procedures before each examination to ensure they applied the same standard.

Data Analysis (Method)

In order to analyse the data and look for any computer effects, students' scores were broken down into several components as described below.

Student's Score

We will define a student's score S in a particular test in terms of i) an overall true mean μ for the test, ii) his individual true offset θ from that true mean, iii) an additive effect ν from the examiner and iv), a residual term R.

μ can be considered as the overall mean score obtained in the test if an entire population of students repeatedly sat for the same test. We cannot estimate (nor do we need to be able to estimate) this figure from our sample.

θ can be considered as the *true* mean score of a student who repeatedly sat for the same test under the same conditions, less the true mean score μ. It follows that the mean θ across *all* students is zero. However, in any sample, the mean θ is generally not zero. While the θ figures cannot be accurately estimated from the sample, the values are not required since analysis is independent of them.

ν is a measure of leniency $(\nu > 0)$ or severity $(\nu < 0)$ on the examiner's part. We assume that, for a given examiner, ν is the same for all students. For example, if he is lenient with a less-able student, we assume he is lenient by the same degree with a more able student. While an examiner may show random fluctuations in his leniency or severity, the mean of those fluctuations is ν and any random variation can be absorbed into the residual term R described below.

For all intents and purposes, the residual term R is a random term which we can arbitrarily define to have mean 0 for each student. In repeated tests of the same student we would obtain a set of independent residuals which have a certain variance σ^2. We assume that this variance is the same for each student and that $R's$ are uncorrelated across student abilities. This term includes variation from the examiners as well as from practice effects (see below).

Thus, if n students sat for the same test with the same team of examiners and S_i denotes the score for student i, we can write $S_i = \mu + \theta_i + \nu + R_i$.

Now suppose the students sit for two different tests, which are respectively examined by two different teams. Denoting the two test means by μ_1 and μ_2 and the two team effects by ν_1 and ν_2 we can write,

$S_{ij} = \mu_j + \theta_i + \nu_j + R_{ij}$ where $j = 1$ for the first test and team, and $j = 2$ for the second test and team.

Further, if a test is administered on a computer, rather than on paper, another additive effect κ is assumed present. κ is not considered a constant but rather, it may be a function of an individual's gender or ability and so on. κ can be interpreted as the number of marks added $(\kappa > 0)$ or subtracted $(\kappa < 0)$ from a student's score when he reads the text from computer screen rather than a sheet of paper.

Finally, in sitting for test 1, a student may have gained some practice for test 2. While the amount of practice may vary from one student to the next, we shall denote the average practice amount by p and any fluctuations from p can be absorbed into the residual term.

The student's scores - S for group I and S' for group II - may thus be written as follows:

Group	Text 1/Team 1, score for ith student	Text 2/Team 2, score for ith student
I	$S_{i1} = \mu_1 + \theta_i + \nu_1 + \kappa + R_{i1}$	$S_{i2} = \mu_2 + \theta_i + \nu_2 + R_{i2} + p$
II	$S'_{i1} = \mu_1 + \theta'_i + \nu_1 + R'_{i1}$	$S'_{i1} = \mu_1 + \theta'_i + \nu_1 + \kappa + R'_{i2} + p$

Subtracting score 1 from score 2 for each student, we obtain a set of differences D:

Group	Score 2 minus Score 1 for ith student
I	$D_i = (\mu_2 - \mu_1 + \nu_2 - \nu_1 + p) + (R_{i2} - R_{i1}) - \kappa$
II	$D'_i = (\mu_2 - \mu_1 + \nu_2 - \nu_1 + p) + (R'_{i2} - R'_{i1}) + \kappa$

Note that the first bracketed term is a constant which we can denote by c and is common to both groups. The second bracketed term in either group is a random term which we shall denote by Z and has expectation zero and, using the assumption of independent residuals, has variance $2\sigma^2$.

Group	Score 2 minus Score 1 for ith student
I	$D_i = c - \kappa + Z_i$
II	$D'_i = c + \kappa + Z'_i$

For a constant κ, an unbiased estimate of its value can be obtained by subtracting the mean difference for Group I from the mean difference for Group II and halving the result (the c terms cancel and the Z terms have expectation 0). However, since we wish to see if the computer effect varies with student background, it is better to do the following:

We define a dichotomous variable G which takes on the value -1 for Group I students and +1 for Group II students. Thus the differences D_{ij} for the ith person in the jth group can be written in just one equation:

$$D_{ij} = c + G_j \times \kappa + Z_{ij}, \text{ where } j = 1,2 \text{ and } G_1 = -1, G_2 = 1$$

Since the values for D_{ij} can be observed and G_j is known, estimates for c, κ and $2\sigma^2$ can thus be obtained by linear regression with G as explanatory variable and D as dependent variable. In this case the estimated coefficient of G is that of the computer effect, the observed intercept is an estimate of c and the residual variance is an estimate of $2\sigma^2$. Checking for interaction of the computer effect with background variables can also be done using regression. For example, suppose a dichotomous variable F is defined which takes on the value 1 for girls and 0 for boys. In such a case the explanatory variables should be G and G*F – the coefficient for G indicates the computer effect for boys and that for G*F indicates the difference in computer effect between boys and girls – the effect on girls is simply the sum of the two coefficients.

More than two teams

The large number of students involved in the study necessitated the use of several teams of examiners. For example, in just one of the schools at the Express level there were over 180 students to be examined in a single day. Typically an oral examiner would examine a maximum of 15 candidates in one session, though this would be for the entire oral paper including Reading Aloud, Picture Discussion and Conversation. Since only one component – Reading Aloud – was to be examined for this study, we could allow for 45 candidates to be examined by an examiner in a session.

Each of the groups I and II were therefore divided into 4 sub-groups, labelled A, B, C and D.

In total, then, there were 8 groups of students. Group I was divided into groups IA, IB, IC and ID, while Group II was divided into groups IIA, IIB, IIC and IID. The 8 groups were equal in the same sense as described above.

There were also 8 examiner teams - A1, B1, C1 and D1 in place of Team 1 and A2, B2, C2 and D2 in place of Team 2. Teams were paired in that those students who were examined by Team A1 for Text 1 were also examined by Team A2 for Text 2. Similarly, Teams B1 and B2 were paired as were C1/C2 and D1/D2.

In summary:

Groups	Text 1	Text 2
IA, IB, IC, ID	i) 5 minutes preparation on computer. ii) 4 minutes examination on computer by Teams **A1, B1, C1 and D1** respectively	iii) 5 minutes preparation on paper. iv) 4 minutes examination on paper by Teams **A2, B2, C2 and D2** respectively
IIA, IIB, IIC, IID	i) 5 minutes preparation on paper. ii) 4 minutes examination on paper by Teams **A1, B1, C1 and D1** respectively	iii) 5 minutes preparation on computer. iv) 4 minutes examination on computer by Teams **A2, B2, C2 and D2** respectively.

The necessity of multiple teams introduces an additional point of consideration in our analyses. For example, while differences between team *pairs* such as between A1 and A2, or between B1 and B2 and so on, do not affect our analyses (as the design removes differences which may exist), differences between those differences *does* make a difference. For example, if the difference between A1 and A2 is *different from* the difference between B1 and B2 then all calculations must be based on those differences. Thus the term c used in the equation $D_{ij} = c + G_j \times \kappa + Z_{ij}$ is no longer a constant but one of four values depending upon the subgroup A, B, C or D to which the student belongs. In a regression analysis, three additional explanatory variables, A, B and C, say, which take on the value 1 if a student belongs to that subgroup, but 0 otherwise, must be included. Thus a student from sub-group IA or IIA would have values A=1, B=0 and C=0, whereas a student belonging to sub-group ID or IID would have values A=0, B=0 and C=0. Interactions between G and any of A or B or C are equivalent to interactions between examiners and computers and are not considered here.

The Findings and Discussion

i) Overall simple statistics

Course	Group	#[3]	Mean Test 1	Mean Test 2	Difference (T2-T1)[4]
Express	I	137	8.54 (1.87, 0.16[5])	8.69 (1.79, 0.15)	0.15 (1.53, 0.13)
	II	135	8.60 (1.74, 0.15)	9.06 (1.95, 0.17)	0.46 (1.44, 0.12)
Normal Academic	I	103	8.17 (1.54, 0.15)	8.31 (1.72, 0.17)	0.15 (1.42, 0.14)
	II	93	8.28 (1.81, 0.19)	8.52 (1.95, 0.20)	0.24 (1.69, 0.18)

[3] For the express course, 4 students from Group I and 9 from group II were outliers and not included in the analysis. For the Normal academic course, the numbers of outliers was 8 and 5 respectively.
[4] The difference figures are for information only. They cannot be used in any direct way of calculating an overall computer effect since multiple teams of examiners were used in the study
[5] Figures in brackets indicate standard deviation and standard error of measurement

ii) Constant Computer effect.

When modelling for a constant computer effect – that is, one in which the computer affects all students in the same way – values of 0.12 ± 0.16 for the Express course and 0.07 ± 0.19 for the Normal Academic course were observed. The value 0.12 for the Express course means that on average, 0.12 marks were added to students' results because of using a computer – note that this figure is independent of test difference, examiner differences and practice. However, since the confidence interval 0.12 ± 0.16[6] contains the number 0, the observed effect is not significant. Similarly, the observed effect in the Normal Academic course is not significant.

iii) Computer effect by gender.

For the Express course, the observed computer effect for boys was 0.14 ± 0.23 (not significant), and for girls 0.11 ± 0.22 (not significant). In particular, the 0.03 difference between the observed effects has an error term of ±0.32 and the difference between the girls and boys is therefore not statistically significant.

A similar conclusion is drawn for the Normal Academic course where the boys' observed (non-significant) computer effect of 0.08 ± 0.24 is statistically no different from the girls' observed (non-significant) computer effect of 0.05 ± 0.30. (The difference is 0.04 ± 0.39.)

iv) Computer effect in terms of Socio Economic Status

When the SES coefficient was built into the model, a 0.17 ± 0.28 change in mark for every one standard deviation change in SES was observed for the Express course students. The corresponding value for the Normal course students was 0.06 ± 0.38. The results are not significant in either case.

v) Computer effect by Ethnic group.

When analysing the data with ethnic group as an explanatory variable, no significant computer effect was detected in either course. For the Express Course students, the observed computer effects were 0.10 ± 0.22, 0.21 ± 0.28, -0.04 ± 0.5 and 0.12 ± 1.09 (for Chinese, Malay, Indian and Others). None or these results are significantly different from zero. Furthermore, there is no significant difference among the four ethnic groups. The corresponding figures for the Normal course were: 0.22 ± 0.31, 0.00 ± 0.26, -0.15 ± 0.63 and 0.09 ± 1.57.

Implications

Although the study has shown that there is no significant difference in the performance of the students whether they read off a printed page or a computer screen, the implications of implementing a similar computer-based test nationwide involving high stakes examinations are several.

System-readiness

A system must first be in place to support the delivery, the conduct and the assessment of candidates without any compromises to the integrity of the examination.

---

[6] All confidence intervals in this paper are two standard errors wide. Equivalently, they are 96% confidence intervals.

Teacher-readiness
While putting in place a solid infrastructure to support teaching, learning and assessing in the classroom is imperative, the teachers and examiners must be familiar and comfortable in using both the hardware and software as tools of instruction and assessment. Teachers and examiners must be equipped with the necessary skills and knowledge to use computers in their classroom and examination rooms. They should be able to instruct their students using computers and also be able to troubleshoot and rectify simple technical glitches that their students might encounter in a teaching and testing situation in their classrooms. Without such abilities, the teaching and testing will be seriously hampered and for large scale testing this would indeed be serious.

Student-readiness
In helping students learn using computers, MOE has "woven into the *EL Syllabus2010"* when it revised recently its English Language teaching syllabus, the "use of information and communication technology (ICT)" as well as ideas on "Social and Emotional Learning (SEL)" and "cyberwellness".

The introduction of SEL and cyberwellness shows MOE's concern for students to *be discerning and responsible users of ICT, Blogs, Facebook, (and) Twitter*.

Changes in IT Environment; Changes in Skills Needed
While the Ministry of Education's *ICT Masterplan* aimed at preparing the system, the teachers, the examiners and the students to acquire the essential skills for the 21 Century, there is no guarantee that all will be well - for one, the IT environment is very fluid and change comes about very quickly and there will always be a need to adapt to these changes through re-training and re-learning to acquire new sets of skills.

Conclusion
The findings of this study show that the mode of delivery (computer or paper) makes no difference to Grade 10 students' performance in the reading aloud component of the English Language oral examination. As the nature of this study is exploratory and concerns only with one component of an examination, it may be worthwhile to carry out other studies that might inform us if there are other factors that might significantly interfere with student performance in an examination when computer is used. Other possible areas for study might be how other language skills such as listening, writing and speaking might be assessed using computers and if this mode of assessment might have an impact on students' performance.

References
Iswaran S, (2010) Senior Minister of State, Ministry of Education, Speech made at the International Conference on Teaching and Learning with Technology on 4 March 2010, Accessed 30 June 2010 from http://www.moe.gov.sg/media/speeches/2010/03/04/speech-by-mr-s-iswaran-at-ictlt-2010.php

English Language Syllabus 2010 Primary & Secondary (Express/ Normal [Academic]), Ministry of Education, Singapore, October 2008. Accessed June 21, 2010 from http://www.moe.gov.sg/education/syllabuses/languages-and-literature/files/english-primary-secondary-express-normal-academic.pdf