PEOPLE: International Journal of Social Sciences
ISSN 2454-5899

Anne C. Ihata, 2019

Volume 5 Issue 3, pp.215-223

Date of Publication: 06th December 2019

DOI: https://doi.org/10.20319/pijss.2019.53.215223

This Paper can be cited as: Ihata, A. C. (2019). Comparative Study on Reliability between Online and Paper-based Versions of a Test for Reading in English as a Foreign Language. PEOPLE: International Journal of Social Sciences, 5(3), 215-223.

COMPARATIVE STUDY ON RELIABILITY BETWEEN ONLINE AND PAPER-BASED VERSIONS OF A TEST FOR READING IN ENGLISH AS A FOREIGN LANGUAGE

Anne C. Ihata
Musashino University, Tokyo, Japan
aihata@musashino-u.ac.jp

Abstract

The aim of this study is to examine how reliable the results of the internet-based version of a reading test are, with a view to replacing paper tests with the online versions. This is of increasing importance as universities focus on improving efficiency and supporting SDGs by going paperless. The study was also suggested by the need to test reading comprehension of larger numbers of students across the university and deliver meaningful results on which to base intensive programs of instruction quickly. The Extensive Reading Foundation's online reading test and the (now discontinued) Edinburgh Project for Extensive Reading's placement test (paper-based), were administered to university students under controlled conditions, and the data was analyzed for possible relationships. An initial one-way ANOVA analysis of the results suggested little evidence of a relationship between online and paper-based test scores. However, further analysis using other measures found evidence of interaction between them, and a second ANOVA analysis, only of scores for students who had completed all versions of the test found a significant relationship. Familiarity with both versions of the test was considered as a possible factor. Although this is only a small-scale study, the findings help to support the argument for adopting the online version of the test, with its various potential benefits to schools and educators.

Keywords

EFL, Reading Comprehension, Assessment, Internet-/Paper-Based Tests

1. Introduction

Like many organizations nowadays, schools and universities are under pressure to economize by reducing waste, especially in areas where this also helps to promote Sustainable Development Goals (SDGs). For any sizable enterprise, one of the most obvious targets for saving money and supporting environmental protection is the consumption of paper. In schools this is a crucial resource that we need to conserve or eliminate by finding alternatives. This includes testing as well as regular classes, particularly large-scale tests. Fortunately, with the increasing use of computers and online resources, this is becoming more achievable all the time, and most standard international tests of language skills have online versions which are widely used now.

Although there is some evidence that students may perform better on paper tests than on computer-based versions (Herold, 2016; Backes & Cowan, 2018; for example) or vice versa (White et al., 2015; for example), this is not immediately relevant to the present case, as scores were essentially used to assess students' individual progress in each term of the course, rather than their performance relative to a fixed scale, and paper test scores and online scores were compared to the same mode for this purpose. Of course, the students were also given their results in terms of their achievement on the standard measure in each case, but improvement was emphasized over actual score, which was also given weight only in cases where it was either very low or very high. A comparable study in the field of Informatics at the University of Rijeka in Croatia found that there was no significant difference in median values of the results achieved during online tests and traditional paper-based tests (Candrlic, Asenbrener-Katic & Dlab, 2014). In other words, the evidence in general is still somewhat mixed, although the latter research seems most similar in terms of age and literacy (traditional and computer) levels of participants to the situation in the present case.

The university where the present study took place is firmly committed to integrating SDGs into the curriculum and practices throughout the institution, and at the same time enhancing the learning experience and outcomes for its students as much as possible. Naturally, another important target is reducing costs. In this context, testing and assessment need to be almost constantly reevaluated to streamline processes, ensure the best use of resources, and deliver useful results. Useful here means both to the students and instructors in understanding their own capabilities and areas of weakness for future effort, and also to program coordinators and administrators in determining the success of a program and where it could be strengthened or restructured. Large-scale standard tests (IELTS, TOEIC and TOEFL) are generally used to measure language skills, and these are now mostly computer-based (CBT) or internet-Based (iBT), for convenience in testing large numbers and in obtaining results quickly. They naturally have sections devoted to the testing of reading ability. However, in Japan, the most commonly used test to date is the TOEIC, which focuses on the reading of non-fiction texts mostly with a somewhat limited business orientation, which may not require the full range of skills and strategies that a mixture of non-fiction, expository texts from various fields, and obviously fictional passages might need – the kind of variety that people are likely to encounter in daily life.

It is against this background that students in an EFL reading class were chosen to examine the correlation between the online and paper-based versions of a well-known standard test of reading skills (the progress/placement tests offered by the Extensive Reading Foundation and the Edinburgh Project on Extensive Reading, respectively), based on passages from various works of fiction and some non-fiction texts, as students would be expected to become familiar with through their required extensive reading out of class and their textbook (*Inside Reading 2,* 2nd Ed., Zwier, 2012).

2. Research Issues

Here the key research issue is not the essential validity of either of the tests used as reliable measures of reading ability in English as a foreign language, since they are well-established means of assessment (See Walker, 1997; Azmuddin et al., 2014 and Herbert, 2016, for example). It should, however, be noted that it is the EPER Progress/Placement Tests (A and B) that were used, not the level-specific comprehension tests which have been criticized as not functioning well as assessment tools in a Japanese university context (Yoshizawa, 2014).

The question this study really sought to address was whether or not there would be any significant correlation in values of the results achieved during online tests and traditional paper-based tests taken by the same subjects, which could offer support for using the online tests for the various advantages they offer over paper tests. Not least among these advantages is the instant scoring in test time – a welcome bonus for busy teachers.

3. Method

The students in a single EFL Reading class at university level (Musashino University in Tokyo) were selected as subjects for this research, despite the relatively small size of the group (20 students originally), largely because it was possible to monitor their performance over an entire academic year. Thanks to the university's four term system, this allowed for the paper test and online test to be used twice each, in alternate terms, during the 2018-2019 academic year. All subjects were juniors in the Department of Global Communication, and assessed as having English ability equivalent to CEFR B-1 or higher (measured by standard proficiency tests such as TOEIC). The majority of subjects (13) were Japanese, 5 were Chinese, one Mongolian and one Korean (actually born and brought up for ten years in the US). All the non-Japanese subjects participated in the course for the whole year, except one of the Chinese students who only participated in the latter two terms. Since all subjects took the same tests, two paper-based and two online, no-one should have been particularly benefitted or disadvantaged by the test mode over the year. The student who only attended classes for two terms was one of the most improved in the final test.

The tests do differ in format. The EPER test is a cloze type, with blank spaces for one-word answers, and its scoring is quite severe in that no alternative responses (except where several are listed) and no misspellings are accepted. This rule was slightly modified to allow for minor spelling errors, where the intended word was in no doubt and would have been understandable to anyone not familiar with these subjects. The test consists of twelve (Version A) or thirteen (Version B) short passages which gradually increase in difficulty. The Extensive Reading Foundation online test has a set of True/False questions after each reading passage. The level of difficulty appropriate for the individual test taker is assessed approximately at the beginning of the test. There are usually three passages in each test, longer than those of the EPER test, and reading speed as well as accuracy of answers affects the final score.

Tests were used as end-of term final assessments and one ninety-minute class period was allowed in each case, although those who finished early were allowed to leave, and, perhaps unfortunately, no record was kept of the length of time taken by each subject. Paper tests were scored by the author and one assistant, with careful discussion of any possibly acceptable variations in spelling to ensure a consistent standard.

Results were tabulated and subjected to statistical analysis (using SPSS Version 23) to examine them for possible relationships. Numbers of test takers varied, although there were 19 subjects taking both paper-based tests and 20 taking the second online test but only 12 in the case of the first online test, largely due to timetable clashes with other necessary courses only available to seniors in the second term. This rather low population for the one test may, of course, have influenced the results to some extent.

4. Results and Discussion

The following table (Table 1) contains the basic descriptive statistics for the four test administrations. The only points that should perhaps be noted here are that, although the standard deviations for paper tests seem large compared to those for the online versions, this can be accounted for by the larger range of possible scores than by any statistically relevant greater variation in the results.

Table 1: *Descriptive Statistics*

Test Mode & Version	N	Minimum	Maximum	Mean	Std. Deviation
Online1 (Term 2)	12	4.00	14.90	9.3583	3.15896
Online2 (Term 4)	20	3.10	17.40	9.2850	3.31889
Paper1 (Term 1)	19	18.00	64.00	32.4211	10.93067
Paper2 (Term 3)	19	29.00	77.00	40.0526	11.33566

In terms of reflecting an improvement in students' reading skills over time, this is observable in the case of paper-based results, which were statistically significant at $p \leq 0.05$, but less obvious in the online results, presumably owing to the increase in number of participants in the second test, although even here the maximum score is a good deal higher. It is, of course, possible that the production component (spelling) may account for some of the variability in the paper-based scores. There is no data available in this case to examine whether lack of familiarity with the test format/layout would equally affect subjects in both modes, or only in the computer-based mode. However, the time devoted to explaining what was required and redirecting subjects who encountered problems during their first experience with the ERF test suggests that this is an aspect of online testing that definitely needs to be taken into account at the initial planning stages for a switch in test modes to be realized efficiently. (This is a point that is emphasized by various researchers. Arney (2015), Boevé et al. (2015), and Graham (2016) refer to the importance of ensuring that test-takers are familiar with the computer mode when switching from paper-based to online mode in standard tests. Jüngling et al. (2018) were also concerned that potential students' ability to deal with technical aspects of tasks needed to be included when designing a web-based tool to assess their suitability for a program of study.) The question types, True/False vs. cloze passages, are both commonly used in tests in Japanese schools and universities and would have posed no particular problems for anyone.

We come then to the key issue, the question of whether or not there is a close relationship between the results obtained in the online and paper-based modes. This was examined using a one-way ANOVA and Pearson and Kendall correlation coefficients. The first may have been at least partly influenced by the very small number of participants available for comparison in the case of the first set of online results and proved largely inconclusive, with only a very weak relationship indicated between the initial paper test and the latter online version (significant at $p \leq 0.092$, F = 3.398).

Table 2: *Correlations: Online and Paper-based Test Modes*

Test Mode & Version		**Online 1**	**Online 2**	**Paper 1**	**Paper 2**
Online 1	**Pearson Correlation**	1	.780**	**.767****	**.714***
	Sig. (2-tailed)		.003	.006	.014
	N	12	12	11	11
	Kendall's tau b Correlation	1.000	.595**	**.537***	.434
	Sig. (2-tailed)	.	.007	.023	.069
	N	12	12	11	11
Online 2	**Pearson Correlation**	.780**	1	**.631****	**.666****
	Sig. (2-tailed)	.003		.004	.002
	N	12	20	19	19
	Kendall's tau b Correlation	.595**	1.000	**.343***	**.351***
	Sig. (2-tailed)	.007	.	.045	.043
	N	12	20	19	19
Paper 1	**Pearson Correlation**	**.767****	**.631****	1	.921**
	Sig. (2-tailed)	.006	.004		.000
	N	11	19	19	19
	Kendall's tau b Correlation	**.537***	**.343***	1.000	.692**
	Sig. (2-tailed)	.023	.045	.	.000
	N	11	19	19	19
Paper 2	**Pearson Correlation**	**.714***	**.666****	.921**	1
	Sig. (2-tailed)	.014	.002	.000	
	N	11	19	19	19
	Kendall's tau b Correlation	.434	**.351***	.692**	1.000
	Sig. (2-tailed)	.069	.043	.000	.
	N	11	19	19	19

** Correlation is significant at the 0.01 level (2-tailed).$_{**}$

Correlation is significant at the 0.05 level (2-tailed).

Table 2 above shows the results for the correlation measures, namely Pearson's r and Kendall's tau b. Figures in bold indicate correlations which are particularly relevant to the present question. The Pearson correlation coefficient indicates particularly strong correspondence between Paper 1 and both online versions (r = 0.767, $p \leq 0.006$ and r = 0.631, $p \leq 0.004$, respectively), and for Paper 2 and the second online test (r = 0.666, $p \leq 0.002$), whereas a weaker but still significant relationship is indicated for Paper 2 and the first online results (r = 0.714, $p \leq 0.014$). These relationships are supported by the Kendall's coefficient scores, although these are all at a lower level of significance than suggested by Pearson's coefficient.

The results in general, therefore, appear to provide evidence to support the interchangeability of paper-based and computer-based test modes using these particular tests. This echoes the findings in the Croatian study by Candrlic et al. (2014), mentioned earlier (See Introduction), although this was in a completely different field.

5. Conclusion

In conclusion, this study compared the results of online and paper-based versions of a reading test in English as a foreign language for a relatively small group of students (20 in all) in a Japanese university on two tests presented as being equivalent measures of reading skills in English as a foreign language: the online Extensive Reading Foundation's Placement Test and the paper-based Edinburgh Project on Extensive Reading's Placement Test (Versions A and B). Results of the study indicate that either of the two modes may be reliably used to assess ability, without undue prejudice to any of the test takers. However, it is suggested that, where the computer-based test is to be used for the first time, it is advisable to allow time and opportunity for students to adequately familiarize themselves with the format of the test and what is expected of them (Graham, 2016; Lam, 2018).

Overall, the findings are quite encouraging, since they show that computer-based/online testing and paper-based testing are equivalent in measuring reading skills in an EFL setting, and the former offers various benefits, including avoiding paper waste (important in these days of SDG-conscious institutions), ease and speed of marking, and lower costs. At the same time, it is worth noting that the EPER test's incorporation of an element of production may provide educators with a more holistic view of what students are capable of. The ERF, however, does assess the reading speed component, which may be a useful aspect to explore in the case of the paper-based test in future research.

References

Azmuddin, R., Ali, Z., Ngah, E., Mohd Tamili, L. & Mohd Ruslim, N. (2014). Extensive Reading Using Graded Readers. *International Journal of Research in Social Sciences 3*(8): 109-113.

Backes, B. & Cowan, J. (2018, April). Is the Pen Mightier Than the Keyboard? The Effect of Online Testing on Measured Student Achievement. *CALDER Working Paper No. 190*. National Center for Analysis of Longitudinal Data in Educational Research (American Institutes for Research). Retrieved from https://caldercenter.org/sites/default/files/WP%20190.pdf?platform=hootsuite

Boevé, A. J., Meijer, R. R., Albers, C. J., Beetsma, Y. & Bosker, R. J. (2015). Introducing Computer-Based Testing in High-Stakes Exams in Higher Education: Results of a Field Experiment. *PloS one*, *10*(12), e0143616. https://doi.org/10.1371/journal.pone.0143616

Candrlic, S., Asenbrener Katic, M. & Holenko Dlab, M. (2014). Online vs. Paper-Based Testing: A Comparison of Test Results. *Proceedings of the 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, 775-780. https://doi.org/10.1109/MIPRO.2014.6859649

Edinburgh Project on Extensive Reading: Placement Tests A & B. (1994). Institute for Applied Languages Studies, University of Edinburgh.

Extensive Reading Foundation Placement Test. (n.d.). Retrieved from https://erfpt.ealps.shinshu-u.ac.jp/

Graham, S. (2016). Here's How the Method of Testing Can Change Student Scores. *The Conversation*. Retrieved from https://theconversation.com/heres-how-the-method-of-testing-can-change-student-scores-54992

Herbert, H. (2016, March 16). Extensive Reading Level Placement: Determining Japanese College Students'Appropriate Starting Levels. *Language and Culture: The Journal of the Institute for Language and Culture of Konan University, 20*:143-156. doi 10.14990/00001722

Herold, B. (2016, February 3). PARCC Scores Lower for Students Who Took Exams on Computers. *Education Week, 35* (20): 1, 11. (Print version published February 10, 2016).

Retrieved from https://www.edweek.org/ew/articles/2016/02/03/parcc-scores-lower-on-computer.html?cmp=SOC-SHR-TW

Jüngling, S., Telesko, R., & Reber, A. (2018). Checking the Student's Aptitude for a Bachelor Program: Experiences with a Web-based Tool. *PUPIL: International Journal of Teaching, Education and Learning*, *2*(2). Retrieved from https://grdspublishing.org/index.php/PUPIL/article/view/1505 https://doi.org/10.20319/pijtel.2018.22.149169

Lam, J. (2018). The Pedagogy-driven, Learner-centred, Objective-oriented and Technology-enabled (PLOT) Online Learning Model. *PUPIL: International Journal of Teaching, Education and Learning*, *2*(2). Retrieved from https://grdspublishing.org/index.php/PUPIL/article/view/1432 https://doi.org/10.20319/pijtel.2018.22.6680

SPSS Version 23[Software]. (1989, 2015). IBM Corp.

The Sustainable Development Agenda. (.n.d.). Retrieved from https://www.un.org/sustainable development/development-agenda/

Walker, C. (1997). A Self Access Extensive Reading Project Using Graded Readers (with particular reference to students of English for academic purposes). *Reading in a Foreign Language 11* (1): 121-149. Retrieved from https://nflrc.hawaii.edu/rfl/PastIssues/rfl111walker.pdf

White, S., Kim, Y. Y., Chen, J., & Liu, F. (2015). Performance of Fourth-grade Students in the 2012 NAEP Computer-based Writing Pilot Assessment: Scores, Text Length, and Use of Editing Tools (No. NCES 2015-119). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Retrieved from https://nces.ed.gov/nationsreportcard/subject/writing/pdf/2015119.pdf

Yoshizawa, K. (2014, March). Edinburgh Project on Extensive Reading (EPER) Reading Comprehension Tests: Scoring and Setting Cutoff Scores. *Kansai Daigaku Gaikokugo Gakubu Kiyo*, *10*: 33-43. Retrieved from https://www.kansai-u.ac.jp/fl/publication/pdf_department/10/02yoshizawa.pdf

Zwier, L. J. (2012). *Inside Reading 2* (2nd Ed.). Oxford University Press.