Primary Weaknesses of unsupervised English
Language e-tests with Descriptive and Robust
Inferential Statistics

An unsupervised online language test is inexpensive, time-saving and time-
flexible for administrators and test-takers. However, previous studies
revealed that such test cannot be used for placement. The current research
aims at finding out the reasons underlying this conclusion. The data was
collected by administering a test to 2,892 freshmen at a university in the
province of Aceh, Indonesia. The data was analyzed quantitatively to find
out irregularities in the students' answers, scores, time, and duration of
completing the test. Quantitative analysis was also used to detect the
possibility of cheating. Based on the research results it was found that there
were some primary weaknesses of an unsupervised online language test, one
of which was a tendency by students to skip difficult questions. There were
79% of the participants who skipped at least one question, and 4% skipped
at least 20 questions out of 140 in total. The second weakness was
premature completion, where 28% of the participants finished the test too
erly. Finally, an advanced statistical analysis based on Item Response
Theory revealed that students copied answers from others. Therefore, should
an unsupervised online language test need to be conducted, measures have
to be taken to minimize the weaknesses above.

Keywords: academic dishonesty; low-stakes test; online language test;
placement test

Introduction

Mass language placement tests require much energy and resources.
Therefore, many educational institutions, as surveyed by Latchem and Jung
(2009) and researchers delivered no-stakes and low stakes English proficiency
test online, and students complete the test without supervision. Not only is this
method of test delivery very convenient for test administrators, but it is also
convenient for students. They can complete the test anytime they prefer, and
they can also pause the test when they need to.

In Indonesia, this system has been implemented by the Ministry of
Education since 2017. The program is an online professional development for
in-service teachers. The training includes reading materials, tasks, and tests
which determine the completion of the programs, as well as providing teachers
with an additional monthly incentive after completion of the program. The
government could save a considerable amount of its budget by conducting this
training online. However, Mustafa and Sofyan (2018) compared the scores that
the students obtained in this type of test and those in a traditional paper
delivered supervised test and found that both scores were significantly
different. This suggests that the scores in the unsupervised online tests were
unreliable. The result of this research indicates that online testing, which can be a part of the assessment in online training, has some weaknesses.

Although a research study by Cojocariu, Lazar, and Lazar (2016) has identified the strengths and weaknesses of e-learning educational services, those same weaknesses highlighted in the assessment were not well-addressed. Latchem and Jung (2009, p. 128) also claimed that assessments conducted online pose some security concerns, and that it becomes more serious when the process itself is accessed instead of the performances. Surprisingly, even for distance learning or in the application of virtual university, Mirzakhani, Ashrafi, and Ashrafzadeh (2010, pp. 33–34) did not list assessment as one of the disadvantages of such an educational system. In addition, instructors’ perspectives on distance learning surveyed by Wunder, Elliott, and England (2013) did not address any concerns on assessment, nor did Ostrow and DiMaria-Ghalili (2005). Also, faculty members did not raise any issue regarding assessment in distance learning (O’Neill, Scott, & Conboy, 2011). However, according to Economides and Roupas (2011, pp. 82–83) security, which includes cheating, “is a crucial issue in tests”. According to the research presented in these studies, there is an urgent need to address the primary weaknesses of an unsupervised online assessment. To reveal these weaknesses, this study gathered data from the unsupervised online test results of almost 3,000 students and analyzed it qualitatively and quantitatively using simple and advanced statistical analyses. The results of the current research are expected to generalize to contexts outside of language testing, and thus be better applied to other online pathways for unsupervised testing in the future.

Literature Review

Language Testing

Language testing has been recognized by many to play a crucial role in the teaching of language (Brown, 2004, p. 16). It has influenced many aspects of school curriculum, particularly lesson plans prepared either by the schools or by teachers (Abidin & Jamil, 2015, p. 3). A language test does not always assess what an examinee already knows, like tests in math or history, but it measures proficiency or performance, i.e. the ability to use and understand the language in a real context (Brown, 2004, p. 3). Some tests are high-stakes such as school exit exams, school and job recruitment tests, and standardized language tests like IELTS and TOEFL (Allen, 2014, p. 23). Medium-stakes tests include classroom final exams, and low or no-stakes tests which are for placement and research study purposes (Domínguez et al., 2019, p. 2; Long, Shin, Gresslin, & Willis, 2018, p. 137).

Theories have suggested procedures for designing a language test (Brown, 2004), however in practice most tests are designed according to how it is used or who designed it. A standardized test should be designed based on years of research (Hall, 2009, p. 321), and revisions should be continuously made to
ensure its reliability and validity (Falvey & Shaw, 2006, p. 8). Educational Testing Service (ETS) has a systematic guideline on every stage of test development, such as a manual for fairness review of assessment (ETS, 2009a). The reliability coefficient is analyzed and reported to the public, such as the TOEFL ITP with a reliability coefficient of 0.96 (ETS, 2009b). Tests made by teachers have been reported to have a lower reliability coefficient, as reported by Quaigrain and Arhin (2017, p. 11), for example 0.77, and this number could be lower in other countries and schools. In addition, individuals or institutions designing non-standardized language tests do not usually report the reliability coefficients of their tests.

**Online Language Testing**

The presence of internet access to a wider-scale of users provides another opportunity to change how English language and other tests can be conducted. The cost and time efficiencies that internet access affords has attracted the attention of many institutions, educational or otherwise, in many different countries (Domínguez et al., 2019). Test industries such as Educational Testing Service (ETS) and Cambridge Assessment English have shifted from paper-based tests to internet-delivered counterparts (ETS, 2010, p. 4; Müller & Daller, 2019, p. 6). However, for high-stakes tests, the supervision is as strict as the previous paper-based versions (Lahib, 2016, p. 36). University professors also have shown a positive perception towards the use of e-testing as long as it is conducted in a language lab and supervision is possible (Fageeh, 2015, p. 55).

An unsupervised internet-delivered test version is now trending (Pearlman, 2009), not only for low-stakes tests but also for medium and high-stakes tests. Final exams taken by students enrolled in a distance learning university that determined whether the students passed or failed the program, are one example of an unsupervised online exam (Piotrowski, Guyette, & King, 2009). It is not always possible to properly proctor such an exam. A study conducted by Chuang (2015) revealed that non-verbal cues such as time delay, head pose, and affective state (confusion) were significantly correlated to cheating on an online delivered exam, yet these online proctoring methods are mostly possible only in a computer lab. Chuang (2015, p. 6) also explains that using online proctoring tools such as ProctorU and Kryterion, requires an excessive amount of time and thus needs an unreasonable amount of money to implement.

**Resistance toward Online Language Testing**

The acceptance rate of online language testing is very high (Fageeh, 2015; Ozden, Erturk, & Sanli, 2004); though some teachers and students still resist it for various considerations. The first reason for a bias towards paper-based exams was because the students simply found it more comfortable to complete a paper-based test (Jamiluddin, Darmawati, & Uke, 2017, p. 142). This consideration is related to computer literacy (Demirci, 2007, p. 33). The
students participating in the survey conducted by Jamiluddin et al. (2017) took an online high-stakes language test as their first experience. The level of anxiety in taking an online exam was lower among students with good computer skills (Jimoh, Kehinde Shittu, & Kola, 2012). In a study conducted in the implementation of e-testing outside language courses found that low performing students had a negative attitude towards the use of this system (Sorensen, 2013, p. 184). This attitude might be influenced by their anxiety and test threat (Cassady & Gridley, 2005, pp. 14–15). In addition to students' resistance, many teachers were also reluctant to use the e-learning system for assessment in the classes. Öz (2014, p. 57) analyzed their concern by distributing a set of questionnaires and found that “frequency of internet usage and level of computer literacy” were the cause for this reluctance. Unfortunately, the students and teachers’ perceptions outlined above only apply to supervised online testing. The literature has not provided any insight into the students’ or teachers’ perceptions of the implementation of unsupervised online testing.

**Issues in Online Language Testing**

As much as it is preferred for its efficiency, Mustafa and Sofyan (2018, p. 286) opposed the idea of using an online language test unless it is properly supervised. Previous research has identified some possible problems in conducting online language tests. Fageeh (2015, p. 54) found that the problems of testing online include lack of facilities, problematic internet connection, and lack of technical support in many educational institutions. Most importantly, Impara and Foster (2015, pp. 108–109) voiced their concern on the possibility for the examinees to cheat by copying. Many students admitted that it is easier to cheat when the exam is delivered online (Peled, Eshet, Barczyk, & Grinautski, 2019, p. 56). Moreover, reports by Lucky, Branham, and Atchison (2019, p. 418) confirmed that students did in fact cheat on online exams. A survey conducted by Piotrowski et al. (2009, p. 6) revealed that students exhibited distorted behaviours towards cheating on an online exam. This result is also strengthened by Shachar and Neumann (2010, p. 324) who compared the grades obtained by students enrolled in distance education and those in traditional education and found that the former has outperformed the latter for the past twenty years.

The previous research results presented above deal with problems in a medium and high-stakes language test. The weaknesses of low or no-stakes tests have not been adequately explored. Therefore, the current research utilizes mixed research methods to find out the primary weaknesses of unsupervised online language tests used for placement.
Methods

This section presents a comprehensive description of how the current research was conducted. It includes the type of data, source of data, data collection procedure, qualitative data analysis, and quantitative data analysis.

Research Design

This research utilized a mixed method, i.e. qualitative and quantitative method. The use of both methods was necessary to highlight as many weaknesses as possible. Descriptive statistics were used for qualitative analysis, and inferential statistics were used for the quantitative counterpart.

Source and Type of Data

The data used for this research was collected from unsupervised online TOEFL tests delivered to third-semester undergraduate students at Syiah Kuala University, Indonesia in 2018. As many as 2,892 students took the test as a requirement for taking the final exam for a General English module. The students had taken the same type of test approximately five months previously, but it was delivered in paper format with supervision. The scores they obtained in their final exams, where the students were supervised under a common language test setting, were used to compare the scores with the unsupervised online test score. However, 16 students had not taken the final exam when the data was collected for this research; therefore, the total sample size for comparison was 2,876. The students have learned English in secondary schools for at least six years, and they had just finished one semester of a TOEFL preparation class at the university. The students are from all majors at the university except English language teaching.

Instrument and Data Collection Procedure

The data was collected by administering an online test delivered using Moodle, a popularly used opensource Course Management System (CMS), which was also used for testing purposes by Sorensen (2013). The test used was based on TOEFL material designed by Educational Testing Service, comprising of 50 items of listening comprehension, 40 items of structure and written expression, and 50 items of reading comprehension. All questions were multiple choice with four options. For the listening comprehension test, a special javascript code was used to ensure that the students could not replay the audio; therefore, the audio was only played once. When delivering the test, we announced that the students were not required to pass any minimum score. The test was considered complete after they took the test for at least two hours. The students were required to complete the test within one week whenever they had the opportunity to do so. Therefore, students did not have to take the test at the same time. After the test, the students could see the number of correct
answers, but each test section was not reported separately. They could not see
either which number was correct or the correct answers for the questions.

**Data Analysis**

The data was analysed for irregularities based on the students’ answers
extracted from the e-learning system using qualitative data analysis, which is
almost similar to the procedure of Grounded Theory research method originally
proposed by Glasser and Strauss (1967). In addition to other irregularities, we
also checked for cheating of any kind by using quantitative procedure. In this
case, we compared the students’ answers for suspected cheating based on Item
Response Theory (IRT).

There are some indices which can be used to determine the similarity, such
as ω, GBT, K, K1, K2, S1, and S2 (Zopluoglu, 2013, p. 94). In this particular
research, we used M4 Similarity Index because it considers both the similarity
of correct answers and that of incorrect answers (Maynes, 2014, p. 55), and it
has been claimed as the response similarity index statistic which “provides the
best approximation for computing the probability of observed similarity”
(Maynes, 2017, p. 59). To avoid type 1 error, for this statistical calculation, we
used a p-value of 0.01 (99%). Therefore, not until the p-value was lower than
0.01 is the evidence for cheating considered. All statistical analyses were
performed using R, open-source software for advanced statistical computing.
For cheating and guessing prediction, CopyDetect package by Zopluoglu
(2013) was used. For more simple, less-time consuming calculations in R, the
students’ responses were converted into dichotomous response, i.e. 1/0, where
1 is for a correct response and 0 for an incorrect or skipped response. For more
detail, we ran the following codes in R.

**Figure 1. R Codes for Suspected Cheating by the Student with ID of e403**

```
94 require(copydetect)
95 data.dich = read.csv(file="cheating.suspected.S40.dich.csv", header = TRUE)
96 data.dich.1 <- as.data.frame(data.dich)
97 y2 <- rep(y[400, "e403"], 1) # Edit here
98 x2 <- as.character(data.dich.1[1:4,])
99 pairs.raw <- cbind(p1, p2)
100 pairs <- matrix(c(as.character(pairs.raw), nrow=500, ncol=2))
101 show(pairs)
102 dich <- similarity1(data = data.dich.1,
103 model = "R",
104 person.id = "ED",
105 item.loc = 3:42,
106 many.pairs = pairs)
107 e403 <- cbind(as.character(dichoutput.manypairs)"suspected copy"),
108 as.character(dichoutput.manypairs)"suspected source"),
109 dichoutput.manypairs%S.pvalue, dichoutput.manypairs%GBT.pvalue,
110 dichoutput.manypairs%944.pvalue)
111 write.csv(e403, file = "output.e403.csv")
```

Furthermore, it is less practical to run the analyses for all students;
therefore, we chose the first 500 students with the highest scores (between 99
and 141) in online test as the comparator because cheating is more likely to
happen among students with high scores. Furthermore, we randomly selected
30 students to compare with the 500 comparators. The selection was performed
using `sample` function without replacement, i.e. `sample(data, size = 30, replace`
The details of all scores (n = 2,892), the comparators (n = 500) and the scores we compared them with (n = 30) are presented in the following figure.

*Figure 2. Scores Used in Analysing Similarity Index*

![Graph showing distribution of scores](image)

**Results**

Since the data was analyzed based on qualitative data analysis using descriptive statistics and quantitative analysis through inferential statistics, this section is divided into two for more systematic data presentation.

**Results from Qualitative Data Analyses**

There are some irregularities which were observed in the students’ raw answers and in the total scores. These irregularities are treated as the weaknesses of the online language test for placement.

**Skipping Questions**

The first weakness of online unsupervised English language test is that the students tend to skip many questions. The following table details the description of students skipping questions.

<table>
<thead>
<tr>
<th>No</th>
<th>Range (%)</th>
<th>ALL</th>
<th></th>
<th></th>
<th>LIST</th>
<th></th>
<th></th>
<th>STR</th>
<th></th>
<th></th>
<th>RDG</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Freq</td>
<td>%</td>
<td></td>
<td>Freq</td>
<td>%</td>
<td></td>
<td>Freq</td>
<td>%</td>
<td></td>
<td>Freq</td>
<td>%</td>
</tr>
<tr>
<td>1</td>
<td>1 - 10</td>
<td>411</td>
<td>76%</td>
<td></td>
<td>193</td>
<td>75%</td>
<td></td>
<td>170</td>
<td>75%</td>
<td></td>
<td>209</td>
<td>65%</td>
</tr>
<tr>
<td>2</td>
<td>11 - 20</td>
<td>29</td>
<td>5%</td>
<td></td>
<td>10</td>
<td>4%</td>
<td></td>
<td>9</td>
<td>4%</td>
<td></td>
<td>22</td>
<td>7%</td>
</tr>
<tr>
<td>3</td>
<td>21 - 30</td>
<td>19</td>
<td>4%</td>
<td></td>
<td>3</td>
<td>1%</td>
<td></td>
<td>1</td>
<td>0%</td>
<td></td>
<td>16</td>
<td>5%</td>
</tr>
<tr>
<td>4</td>
<td>31 - 40</td>
<td>36</td>
<td>7%</td>
<td></td>
<td>10</td>
<td>4%</td>
<td></td>
<td>1</td>
<td>0%</td>
<td></td>
<td>7</td>
<td>2%</td>
</tr>
<tr>
<td>5</td>
<td>41 - 50</td>
<td>11</td>
<td>2%</td>
<td></td>
<td>1</td>
<td>0%</td>
<td></td>
<td>2</td>
<td>1%</td>
<td></td>
<td>9</td>
<td>3%</td>
</tr>
<tr>
<td>6</td>
<td>51 - 60</td>
<td>10</td>
<td>2%</td>
<td></td>
<td>5</td>
<td>2%</td>
<td></td>
<td>8</td>
<td>4%</td>
<td></td>
<td>5</td>
<td>2%</td>
</tr>
<tr>
<td>7</td>
<td>61 - 70</td>
<td>19</td>
<td>4%</td>
<td></td>
<td>1</td>
<td>0%</td>
<td></td>
<td>0</td>
<td>0%</td>
<td></td>
<td>2</td>
<td>1%</td>
</tr>
<tr>
<td>8</td>
<td>71 - 80</td>
<td>7</td>
<td>1%</td>
<td></td>
<td>0</td>
<td>0%</td>
<td></td>
<td>6</td>
<td>3%</td>
<td></td>
<td>9</td>
<td>3%</td>
</tr>
</tbody>
</table>

*Table 1. Number of Students Who Skipped Questions*
Table 1 shows that from 542 students (2.4% of the total sample size) who skipped questions, 411 (79% of them) skipped between 1% and 10% of the questions. When analyzed separately for each test section, students tended to skip reading questions more often than those of other test sections. In general, the number of students skipping fewer questions was lower than those skipping more questions.

Completing the Test Too Quickly

The second weakness of the online language test which was delivered without supervision is that many students finished the test too early. In the instruction, they were told that they must have spent at least two hours to complete the test because the test was designed to be completed in two hours. However, the information in Table 2 revealed that they did not follow the instruction.

Table 2. Duration of Test Completion

<table>
<thead>
<tr>
<th>No</th>
<th>Range in minutes</th>
<th>Freq</th>
<th>%</th>
<th>The duration is:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 - 20</td>
<td>282</td>
<td>10%</td>
<td>too short.</td>
</tr>
<tr>
<td>2</td>
<td>21 - 40</td>
<td>252</td>
<td>9%</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>41 - 60</td>
<td>303</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>61 - 80</td>
<td>333</td>
<td>12%</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>81 - 100</td>
<td>404</td>
<td>14%</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>101 - 120</td>
<td>399</td>
<td>14%</td>
<td>good.</td>
</tr>
<tr>
<td>7</td>
<td>121 - 140</td>
<td>539</td>
<td>19%</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>141 - 170</td>
<td>380</td>
<td>13%</td>
<td>too long.</td>
</tr>
</tbody>
</table>

Table 2 shows that more than 50% of the participants finished the test in less time than what they were instructed. In addition, 29 percent of the participants finished the test even within one hour or less. Although some students spent too much time completing the test, the number was not very alarming (32%).

Low Reattempt Rate

The students were also advised to reattempt the test should their score in the first attempt be not satisfactory. However, very few students retook the test regardless of their scores. Based on the data, only 150 out of 2,892 students (5%) reattempted the test. The number of students who reattempted the test is presented in Table 3.
Table 3 shows that from 150 participants who reattempted the test, only 69% of them had a score of 50% or less in the first attempt. The rest were students having a score higher than 50%, and they were actually not required to reattempt the test.

Unreliable Scores

To find out qualitatively whether the scores that the students obtained represented the students’ English proficiency, the scores were compared to their final exam scores. In the exam, the test used was similar in format and difficulty level to those on the online version. The final exam was delivered in a classroom where the students were strictly supervised. The results are provided in Figure 3 below.

Exam

The figure above illustrates that the students’ scores only decreased up to 58. However, the students whose scores were higher in the online test got the difference between those tests up to 115. In addition, there are three times as
many students whose scores increased as those who obtained lower scores in
the online test.

**Results from Quantitative Data Analysis**

Quantitative analysis was used to detect the presence of cheating and
correlations between some independent findings found in the qualitative
analyses.

Cheating by Copying

Based on Figure 1 above, we selected 500 students with highest score
differences and determined how many of them copied answers from any of 30
randomly-selected friends or sought assistance or even asked them to complete
the test. For that purpose, we used M4-Indexed proposed by Zopluoglu (2013).
The results of the statistical calculation are presented in the following table.
For the reason of space limitation, we only randomly selected ten copiers and
ten sources to be visualized in the table.

**Table 4. Selected P-Values of M4 Index for Copying Detection**

<table>
<thead>
<tr>
<th>ID</th>
<th>e424</th>
<th>e212</th>
<th>e347</th>
<th>e089</th>
<th>e191</th>
<th>e228</th>
<th>e381</th>
<th>e457</th>
<th>e128</th>
<th>e443</th>
</tr>
</thead>
<tbody>
<tr>
<td>e018</td>
<td>0.5195</td>
<td>0.5355</td>
<td>0.5067</td>
<td>0.2868</td>
<td>0.5355</td>
<td>0.5355</td>
<td>0.3449</td>
<td>0.6210</td>
<td>0.7267</td>
<td>0.3006</td>
</tr>
<tr>
<td>e161</td>
<td>0.9075</td>
<td>0.4975</td>
<td>0.4478</td>
<td>0.3214</td>
<td>0.4975</td>
<td>0.4975</td>
<td>0.6173</td>
<td>0.9459</td>
<td>0.6487</td>
<td>0.5769</td>
</tr>
<tr>
<td>e164</td>
<td>0.8965</td>
<td>0.9549</td>
<td>0.1618</td>
<td>0.0529</td>
<td>0.9549</td>
<td>0.9549</td>
<td>0.5789</td>
<td>0.9988</td>
<td>0.9817</td>
<td>0.5377</td>
</tr>
<tr>
<td>e241</td>
<td>0.2500</td>
<td>0.0426</td>
<td>0.4261</td>
<td>0.0244</td>
<td>0.0426</td>
<td>0.0426</td>
<td>0.5958</td>
<td>0.7231</td>
<td>0.0988</td>
<td>0.1254</td>
</tr>
<tr>
<td>e261</td>
<td>0.7174</td>
<td>0.1272</td>
<td>0.7181</td>
<td>0.4414</td>
<td>0.1272</td>
<td>0.1272</td>
<td>0.9407</td>
<td>0.1653</td>
<td>0.2362</td>
<td>0.9062</td>
</tr>
<tr>
<td>e270</td>
<td>0.2335</td>
<td>0.4250</td>
<td>0.4068</td>
<td>0.4489</td>
<td>0.4250</td>
<td>0.4250</td>
<td>0.1225</td>
<td>0.6908</td>
<td>0.5789</td>
<td>0.4528</td>
</tr>
<tr>
<td>e295</td>
<td>0.5041</td>
<td>0.3749</td>
<td>0.1858</td>
<td>0.1222</td>
<td>0.3749</td>
<td>0.3749</td>
<td>0.2178</td>
<td>0.9590</td>
<td>0.1073</td>
<td>0.1168</td>
</tr>
<tr>
<td>e366</td>
<td>0.8676</td>
<td>0.5799</td>
<td>0.1581</td>
<td>0.2803</td>
<td>0.5799</td>
<td>0.5799</td>
<td>0.3945</td>
<td>0.6098</td>
<td>0.7614</td>
<td>0.3493</td>
</tr>
<tr>
<td>e469</td>
<td>0.0784</td>
<td>0.3929</td>
<td>0.7167</td>
<td>0.3742</td>
<td>0.3929</td>
<td>0.3929</td>
<td>0.8049</td>
<td>0.5260</td>
<td>0.1367</td>
<td>0.6451</td>
</tr>
<tr>
<td>e498</td>
<td>0.9385</td>
<td>0.9840</td>
<td>0.9829</td>
<td>0.9969</td>
<td>0.9840</td>
<td>0.9840</td>
<td>0.8266</td>
<td>0.0014</td>
<td>0.2908</td>
<td>0.7540</td>
</tr>
<tr>
<td>MIN</td>
<td>0.0049</td>
<td>0.0034</td>
<td>0.0142</td>
<td>0.0136</td>
<td>0.0034</td>
<td>0.0034</td>
<td>0.0130</td>
<td>0.0000</td>
<td>0.0016</td>
<td>0.0041</td>
</tr>
</tbody>
</table>

The header row represents ten randomly selected students from 30
students who were also randomly selected as the sample for the calculation.
The stub, the very left column, is the ten selected students who were suspected
as copied sources, randomly selected from 500 students. In addition, the body
is the p-values of M4-index. The last row is the minimum p-values from the
data before the ten sources were selected. Based on Table 4, most minimum p-
values were less than 0.005 (p < 0.005). The results from the complete data are
summarized in Table 5 below.
Table 5. Summary of M4-Index P-Values

<table>
<thead>
<tr>
<th></th>
<th>p &lt; 0.001</th>
<th>p &lt; 0.005</th>
<th>p &lt; 0.01</th>
<th>p &lt; 0.05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of persons</td>
<td>7</td>
<td>18</td>
<td>21</td>
<td>30</td>
</tr>
<tr>
<td>Percentage (total n = 30)</td>
<td>21%</td>
<td>60%</td>
<td>70%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Table 5 shows that at the significance level of 0.01 (99% chance of correctness), 70% of the randomly selected students were suspected cheating by copying from others. The percentage is slightly lower at the significance level of 0.005, i.e. 60%.

Skipping vs Duration and Scores

To find out whether the duration of test completion is correlated to the number of questions skipped, a correlation was calculated using Spearman rank correlation because the data was not normally distributed based on the Shapiro-Wilk test. The result of the correlation calculation is presented in Table 6.

Table 6. Results of Correlation Analyses

<table>
<thead>
<tr>
<th>Correlation pairs</th>
<th>S</th>
<th>R</th>
<th>p-value</th>
<th>level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration vs skipping</td>
<td>20745099</td>
<td>0.2225567</td>
<td>0.0000</td>
<td>Low</td>
</tr>
<tr>
<td>Duration vs score</td>
<td>2969073642</td>
<td>0.2634921</td>
<td>0.0000</td>
<td>Low</td>
</tr>
<tr>
<td>Skipping vs score</td>
<td>36347566</td>
<td>-0.3621614</td>
<td>0.0000</td>
<td>Low</td>
</tr>
</tbody>
</table>

Table 6 shows that skipping questions was significantly correlated to how long students took the test (r = .22, p < .000). In addition, the table also shows that the longer time the students spent completing the test, the higher their scores were (r = .26, p < 0.000). However, the correlation coefficients (R) of 22% and 26% are categorized into low-level correlation. The analysis results also show that the skipping questions was negatively correlated to and their scores (r = -0.36, p < 0.000). Better visualization of correlation is provided in Figure 4.

Figure 4. Correlation of Duration Vs Percentage of Questions Skipped and Duration vs Score

The figures above illustrate the correlations between variable pairs separately. The correlation line for each figure represents the best fit for the
distribution of data points. Since the data was categorical data, the correlations do not substitute causation.

Discussion

The current research was based on the fact that "we do not wish to return to computer-based language testing" where a special language lab is required, and some technical assistance and test supervisors are walking around to make sure no one is cheating. However, online language testing, particularly in distance learning, is still in its infancy and much research is required to make it work as it does with traditional paper-delivered, supervised tests (Suvorov & Hegelheimer, 2013, p. 14). Therefore, the objective of this research was to outline the primary weaknesses of such tests in order to advise test administrators to devise a better plan for future application. After analyzing the data qualitatively, several weaknesses were found in the current practice of online unsupervised language testing, i.e. skipping questions, fast test completion (duration), low reattempt rate, and non-representative scores. Further quantitative analyses also revealed that skipping questions was significantly correlated to duration (positive) and score (negative), and duration was also positively correlated to score. A more advanced statistical analysis using Item Response Theory (IRT) with 3PL model showed that, at the significance level of 99%, 70% of 500 students whose scores increased the most cheated by copying.

Our correlation analysis (quantitative) shows that students skipped questions after they spent too much time completing the test. Therefore, we can conclude that they skipped difficult questions. Skipping questions can be disabled using the current e-learning platform, but in this test, it had to be enabled in order to give students some control of which questions they wanted the answer first. However, the students did not return to the questions they skipped as initially expected. Based on the result of correlation analysis, the number of questions they skipped negatively correlates to their score, meaning that the more questions they skipped, the less score they obtained. The ability to skip questions in an e-testing system is similar to what students can do in paper-based tests. In a supervised paper-based test, we found that students skipped much fewer questions. Therefore, we can predict that students would be more willing to answer all questions if the online test was supervised. To avoid this weakness of unsupervised online tests, all questions can be set as "required", and thus the students will not be able to submit their test attempts before answering all questions.

Another weakness of an unsupervised online language testing is a tendency among students to complete the test too quickly. In our study, students were strongly advised to complete the test in no less than two hours. However, our data revealed that 55% of the students completed the test faster than what they were instructed, and 29% of the students finished the test in one hour or faster. These analysis results suggest two possibilities. First, students
might be less serious in taking the test because, in this study, the results of the test did not give them any benefit other than familiarity with the test before they took the real final exam in the next two weeks. In fact, our statistical analysis shows that the students who completed the test faster tended to obtain lower scores than those who took longer to complete the test. This is in line with our initial prediction that the score was not important to them, so they did not put much effort into taking the test. Another possibility is that the students cheated by copying from other students who already completed the test. This is true for students who obtained good scores regardless of short test duration.

Regarding the cheating by copying, which is the most significant weakness of an unsupervised online language test that was found, the data revealed very alarming results. In fact, Impara and Foster (2015) voiced their concern about the possibility of cheating by copying, and the empirical data confirms that it is the case. Based on 30 students who were randomly selected from 500 students who had the highest improved scores, 21 of them (70%) were suspected of cheating by copying according to the advanced statistical analysis using Item Response Theory (IRT) with 3PL model analyses. The possibility of this number to happen by chance is very low since the decision was based on a very low p-value, i.e. 0.01. This cheating is very difficult to prevent in a culture where academic honesty is still low such as in Indonesia, and where plagiarism is also not well understood (Mustafa, 2019). Therefore, much research is required to devise better planning for a cheating-free online language test without supervision. As of now, such a test cannot be fully relied upon until cheating by copying is addressed and resolved by the academic community as a whole. One method to minimize cheating by copying is the use of Safe Exam Browser where students are locked from opening any other application while the exam is underway (Reuter et al., 2019). Questions can also be randomized, so students cannot copy from others although they are completing the exam together. Using an LMS such as Moodle, questions can also be delivered in sequential mode where students cannot return to any previous questions, which can provide better exam security. However, installing Safe Exam Browser on a computer requires some computer skills because it often fails during the installation.

Another significant weakness of an unsupervised online language test is the students’ unreliable scores. This should be our primary concern because a language test is intended to measure the students’ language proficiency. When the scores obtained by the students do not represent their proficiency, the test is essentially rendered moot. The analysis results show that more than 50% of the students obtained higher scores on the online test compared to the paper-based supervised test. The differences are very significant based on the descriptive statistical analysis used in this research or based on an inferential statistical analysis in previous research by Mustafa and Sofyan (2018).

Finally, we found that many students did not answer the questions as they were expected to. In our experiment, we allowed students to reattempt the test if they obtained scores lower than 50%. However, more than 30 percent of students who obtained low score did not reattempt the test. This unexpected
result is more significant for classroom practice more than the testing purpose. If the test is a practice test or an exercise where students are asked to answer the questions in order they can understand a concept that has been taught in class, some measure should be made to ensure that the students complete the test seriously. Mustafa, Raisha, and Mahlil (2021) developed a rubric to assess the quality of student learning process including how they complete a practice exercise. Therefore, should the exercise be delivered online, teachers can grade student learning process as a part of formative assessment using the rubric. A study by Bugbee et al. (2017) found that students completed their work more seriously when it was graded using a rubric.

The results of the current research are intended to make a generalization to wider language testing contexts, and if possible, to content knowledge testing contexts. However, there are some limitations which need to be considered. First, although the analysis, especially for the case of copying, our analysis was only possible for 30 randomly selected students. Analyzing all 2,892 individual students was simply not possible at this stage. If a future analysis is possible to effectively compare all students' responses, such as through permutation technique, the results of the current study can be confidently confirmed. Another limitation deals with the absence of qualitative data. An in-depth anonymous interview with the students can help to make a better, more accurate conclusion, and it will be possible to reveal their reasons for all the results obtained in the current research. The other limitation concerns potential bias in interpreting the data regarding skipped questions and test duration. There is a possibility that students did not intentionally skip questions, but the e-learning platform failed to save their answers due to internet connection problems or software bugs. In addition, the long duration in completion of the test can also translate to a slow internet connection.

**Conclusion and Implications**

The current research was aimed at finding the weaknesses of an unsupervised online language test. Based on 2,892 scores collected from an unsupervised online test delivered through Moodle LMS and 2,876 scores obtained in a supervised paper-delivered test, it was found that there are some weaknesses of an online language test when delivered without supervision. The weaknesses include high tendency to skip questions, short test duration, insincerity which is evident from the students’ hesitation to reattempt the test regardless of their low scores, cheating by copying, and non-representative scores.

Based on the results above, it can be concluded that an unsupervised online language test cannot be used even for low or no-stakes tests. However, it is still possible to use the test for initial screening to avoid mass paper-based testing. For low or no-stakes tests, the test takers who obtained scores higher than the mean score can be given another paper-based test for confirmation. For medium-stakes tests, the test takers who reached the passing grade can be
retested or validated through another paper-based test or interview. An unsupervised online language test may not be avoided at all as long as the results are interpreted with caution. The results of this research have provided insight into how the results of such a test should be interpreted.

References


ETS. (2009b). The TOEFL®ITP reliability table, 1, 573.


Mustafa, F., & Sofyan, H. (2018). Comparing the scores in paper and internet-delivered TOEFL: Can an unsupervised online English language test be used for


