Influence of Duration for Composition and Letter Writing Examination on Performance: An Experimental Design Approach

Molefhe Mogapi¹, Ernest B. Fetogang² & Gaelebale N. Tsheko³

¹Department of Educational Foundations, University of Botswana, Gaborone, Botswana
²³Department of Educational Foundations, University of Botswana, Gaborone, Botswana.

Abstract: Construct irrelevant variance is one of the major threats to generation of valid and reliable scores. Construct irrelevant variance occurs when performance of candidates in an examination is not explained or accounted for by the subject matter assessed but the variance observed in the scores is influenced by other intervening factors. The duration of the examination is one critical element that can have a negative impact of validity of test scores due to the fact that an inappropriately timed examination can severely limit a candidate’s ability to demonstrate what he or she knows understands and can do. Primary School Leaving Examination requires candidates to write composition and letter pieces within a period of 60 minutes. The relevance and appropriateness of the 60 minutes duration has not been determined. There is need therefore to collect empirical evidence to support or challenge the relevance of the 60 minute duration. The current study uses an experimental design to compare performance of learners required to write composition and letter under different experimental conditions. The establishment of an appropriate duration would significantly enhance the construct validity of the examination and subsequent inferences made on the basis of the generated scores.

Key Words: Performance, construct irrelevant variance, time limit, construct validity.

1. Introduction

Psychometric instruments can generally be categorized into two groups; namely speed tests and power tests. A speed test is mainly designed to assess the rate of speed at which the examinee is able to accomplish a given task as well as the level of accuracy demonstrated. The time limit imposed may not allow all candidates to respond to all items. In contrast, power tests contain questions of varying difficulty and the time allocated for the test is enough to allow all candidates to complete the task. However, the speed and power dichotomy may only be useful theoretically because some time restriction is usually imposed to standardize the testing procedures for purposes of grading and selection. Powers and Fowles (1997) touched on this point when they observed that, “…while most standardized tests are intended primarily to reflect test takers’ intellectual “power” (rather than the rate at which they work), virtually all such tests involve some element of speed, however minor” (p. 1). The need to standardize testing conditions has often lead to arbitrary setting of time limits on a variety of assessment instruments. Arbitrarily imposed time limits on examinations that are used to make high stake decisions have been criticized by measurement experts for a variety of reasons.

Firstly, time restriction creates a situation where examinees who are able to write fast are rewarded more than those who are more reflective in their writing styles. A student who reflects on what he or she has written will need more time for planning and review of the work. Secondly, speeded examinations have been criticized because the duration as specified can be a source of disadvantage for certain individuals and subgroups. According to Evans and Reilly (1973), “standardized academic aptitude test have been the subject of persistent criticism from members of certain minority groups who charge that such tests are unfair to members of their groups (p. 173). Flaugher (1970) identified three potential sources of unfairness; these are the content sampled, the way in which the test is administered, and the way in which the scores are used. All of these three points pose a serious threat to test reliability validity. Specifically, unfairness due to the way in which the test is administered may largely be attributable to time limit issues. Enforcing time limit actually means that candidates are being assessed on a totally different construct as “…the imposition of severe time constraints induces an altogether different mode of writing called ‘assessment genre’ …that renders questionable the comparability of the writing skills displayed in timed and untimed contexts” (Lau, 2013, p. 1).

According to Messick (1989), “Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment” (p. 9). The author further states that “In
Up to 1996, grading and eventual certification of primary school leaving examinations candidates was done by means of a norm referenced testing (NRT) procedures where a candidate score was compared to the scores of other candidates in the same group. The procedure worked very well within an environment where primary mission of testing was selection. However, by the late 1980s, NRT lost its relevance as an assessment programme for several reasons; some of the reasons are discussed below. Firstly, content in the new curriculum for Basic Education was stated in explicit terms to show competencies or skills the learner must exhibit at the end of the programme. Basic Education programme laid more emphasis on practical skills, acquisition of problem solving skills and critical thinking skills. Such changes had far reaching implications on national assessment procedures as well as certification.

Secondly, by 1993 the country had almost achieved 100% progression rate from primary to junior secondary education (Government of Botswana, 1999). The relevance of PSLE as selection instrument ceased. Therefore, there was need then to put in place an assessment systematic capable of generating high quality diagnostic information at the national level. Thirdly, NRT was criticised for the apparent weakness in its ability to detect improvements in the system over the years as well as providing high quality feedback information for the teachers in the classroom. Fourthly, a strong criticism was raised against NRT to the effect that it failed to present a comprehensive evaluation of the learner in terms of specific knowledge and abilities as stated in the curriculum.

It was against this background that a diagnostic system was suggested as a method that could best serve the needs of the system at both national, regional, district, school and learner levels. The quotation below succinctly captures the need for reform in the assessment field:

Emergence of Criterion Referenced Testing Procedures:

The main aim of criterion referenced testing system was to accurately profile each learners’ level of proficiency with respect to the content domain. Therefore, assessment was to focus on what the learner knows, understands and can do in relation to the subject matter outlined in the syllabus. According to the Report of the National Commission on Education (1993);

Implementation of a Criterion Reference Testing System:

CRT was first piloted using a geographically representative sample of 41 primary schools. The pilot study was also to create awareness among teachers, head teachers and school inspectorate about the importance of using testing to inform the instructional process. After the pilot exercise was

---

2. Theoretical foundation

Norm Referenced Testing Procedures:
Purpose of the Study

The main objective of the research study is to determine the most suitable duration for composition and letter writing components. Thus, the research question of the study is, “Would composition and letter writing scores differ significantly if the participants are required to write composition and letter under different time limits?”
5. Statement of the Research Hypotheses

To contribute solutions to the problem of this study, the following research hypotheses were tested:

H$_{A1}$: There will be significant difference in the scores of participants required to write composition under different time limits.

H$_{A2}$: There will be significant difference in the scores of participants required to write a letter under different time limits.

H$_{A3}$: The time limit effect will have a significant impact on the correlation between composition and letter writing scores.

6. REVIEW OF RELATED STUDIES

Suh, Kang, Wollack and Kim (2006) correctly observed that there are several reasons why an examinee might not answer all items in a test or complete a written task assigned. The reasons may be due to random or systematic variables. The influence of time limit on performance is systematic because test duration is part of the test instructions that an examinee must follow. The literature review presented in this section documents some of the studies done to determine the relationship between time limit examinee performance.

The New Jersey College Basic Skills Placement Test (NJCBSPT) was developed to assess students’ ability to express themselves in standard English Language. The scores from the test are used to place students in State supported colleges in the United States of America. NJCBSPT is described as a battery of tests in reading, writing and mathematics. The written component of the test is made up of an essay that has to be written in 20 minutes. It became apparent over the years that some students did not actually complete the task assigned. Therefore, there was need to collect empirical evidence to determine the relevance of the 20 minute time limits for students with varying abilities. Livingstone (1987) conducted a quasi-experimental study aimed at providing answers to the question, “What effect will an extra ten minutes of writing time or planning time have on NJCBSPT essay scores”. The research study essentially combined two experiments in one setting; the first experiment compared the original 20 minutes time limit to the proposed 30 minutes duration while the second experiment allowed the researcher to compare the 20 minute time limit to a condition where participants were given 10 minutes planning time and 20 minutes writing time. In both experiments, the control group used the traditional 20 minutes designated for planning. Most significantly, the addition of 10 minutes tended to lower the scores of the low ability students in certain condition. Generally, the study did not produce definitive answers indicating the relationship between time limit and performance.

Educational Testing Service (ETS) has a long history of generating tests designed to evaluate the ability of candidates to communicate effectively in Standard English. Determining reliability and validity of such instruments has become a standard procedure for ETS. The Graduate Record Exam (GRE) is an example per excellence. GRE writing component has a time limit of 40 minutes, the concern for ETS was whether the 40 minutes catered for all students or whether different subgroups were differentially affected by the time allotment. Powers and Fowles (1997) carried out a study to determine the influence of time limit on academic performance; the following research questions were stated:

a) To determine how different time limit affect examinee scores.

b) To investigate the extent to which slow, deliberate test takers may be at a disadvantage under the proposed time limit.

c) To ascertain the relationship of scores under more or less stringent times conditions to several non-test indicators of writing ability.

d) To estimate any joint effect on test performance of time limits and topic disclosure before the test is administered.

A total of 304 participants were recruited and each participant was required to write two essays. One essay was allotted 40 minutes while the second one was given 60 minutes. To test for the significance of topic disclosure, one of the two topics was revealed to candidates before the actual administration. Also participants were required to
indicate whether they regard themselves as ‘slow’, ‘average’, or ‘fast’ writers. A post-test survey instrument was also administered to participants so that they indicate their perceptions relating to the adequacy of the time limit of the two essays. Hierarchical regression statistical procedures were used to analyse the scores. Five independent variables (i.e., time limit, essay disclosure, time limit by disclosure, slowness/quickness, and time limit by slowness/quickness) were added sequentially to determine the magnitude of the effect as well as test for significant interaction between the variables. The result of the analysis showed that the effect of time limit was significant showing that performance significantly improved when time limit was increased from 40 minutes to 60 minutes. Also, participants who regarded themselves as being fast writers performed significantly better than the ‘slow’ and ‘average’ ones (Powers & Fowles, 1997). This essentially shows that no subgroup benefitted more than the other groups. The researchers made a conclusion to the effect that “time limits, even within a relatively limited range (40 vs. 60 minutes), do matter for the direct measure of writing that was investigated here.” (p. 14).

Time limit is an important variable to be studied and understood because limited time has serious negative impact on the ability of individuals to make good decisions. The interest on the relationship between time limit and decision making emanates from concern that limited time has a detrimental effect on decision making (Diederich, 1997; Busemeyer & Diederich, 2002). A study was carried out by Paola and Gioia in 2014 to investigate the effects of time limit on performance. Specifically, the objective of the research study was stated as, “to investigate how and to what extent being exposed to time pressure affects individual performance and whether there is heterogeneity in the ability to handle time pressure” (p. 3). The research study used 220 undergraduate students from the University of Califia in Italy. At the beginning of the study, participants were given an option to choose from two assessment schemes. Under the first assessment scheme, participants were required to write the usual end of semester exam. Those who opted for the second scheme were given two intermediate examinations; one in the middle of the course and the other one at the end of the course. The intermediate test group were randomly assigned to two testing conditions; one mid-term test was to be written within 30 minute time limit whilst the other test was written under condition of no time pressure. The same conditions were also used at the end of the course for the experimental group.

The regression analysis applied on the data indicated a negative but statistically significant relationship between time pressure and performance. In other words, students who operated under time pressure condition obtained an overall grade that was 3.2835 points lower than those operating under no time limit pressure. When other demographic variables were added to the model (such as gender), the general performance remains largely unchanged with the control group outperforming the time limit group. Of particular interest is the statistical significance observed for the female and ability specifications. Further analysis of the data revealed that the observed negative effect of time pressure on performance largely attributable to female performance (Paola & Gioia, 2014). Female students having to perform the task under time pressure obtained an overall grade that was 5 point lower than a female student with no binding completion time. The analysis further indicated that male students working under time pressure did not expect a lower grade while females expected their final grade to go down by three points. As such, Female students suffer from a strong worsening of their performance when operating under time pressure (Paola & Gioia, 2014). Put in other words, the ability of females to think clearly and make good decisions is severely compromised by time pressure.

The California Bar Examination (1980 Version) is basically a selection instrument taken by lawyers from different training institutions in the United States of America. One component of the examination comprises two sections being the Multistate Bar Examination (MBE) and the Essay Examination. The MBE section was made up of 60 multiple choice items covering area such as Constitutional Law, Contracts, Criminal Law, Evidence, and Torts. The essay section on the other hand had four questions assessing knowledge of Business Law and Trial Law. According to Klein (1981), there have been persistent complaints by candidates who took the exam to the effect that time allocated for the examination was too short. Recommendations were made to investigate the validity of these complains. A research study was conducted to determine the extent to which the proposed substantial increase in the time limit for both sections would benefit students.

The participants for the study were 2940 applicants who had applied to write the California Bar Examination in 1981. The applicants were randomly divided into four groups; each group was required to take two separate MBE multiple choice examinations. In the first test, participants were given 55 minutes to answer a set of multiple choice questions and another 90 minutes to answer the second set of multiple choice questions. It come out clearly that the 90 minutes scores are consistently higher than the 55 minutes scores. The average difference between the experimental condition was .87 points indicating that “if applicants were given essentially unlimited time to complete the MBE’s 200 items, the mean raw score on this test would...
increase by about six points” (Klein, p. 8). Further interrogation of the data revealed that 58% of the applicant managed to improve their scores when given an extra time, however, there was no systematic relationship between performance on the test and demographic variables such as gender, race and religion. A post-test survey focusing on the two time limits indicated a preference for extra time by most candidates.

Tsheko and Mogapi (2015) conducted a survey research to document views of teachers concerning the current 60 minutes duration for PSLE composition and letter writing examination. A total of 21 schools and 221 teachers were sampled; the teachers responded to a questionnaire containing five Likert scale type questions. Sixty six percent the teachers were of the view that the current 60 minutes duration is not enough while about 31% of the teachers are of the view that the allocated time is enough. Generally, the views expressed by the teachers indicated preference for an extended time limit because examinees need time to read the stimulus material, brainstorm and use the ideas generated to develop a conceptual map or draft. The draft gives the writer the opportunity to add new ideas, arrange the ideas in a logical and coherent manner before making the final product” (Tsheko & Mogapi, p. 898).

7. METHODOLOGY

The research study follows a true experimental quantitative approach where all the 90 primary school learners were randomly assigned to experimental groups. The independent variable in this case is the duration of the test. The independent variable has three levels being 30 minutes time limit, 60 minutes time limit and 90 minutes time limit. Analysis of variance was used to compare the performance of the three experimental groups at an alpha level of .05%. Composition and letter scores were correlated and tested for significance to determine whether time limit has significant influence of relative standing of each candidate in the norm group.

All current learners in one school doing standard seven in the study were used to avoid cases where some learners are not selected. Most of the learners in standard seven classes are about twelve years and the number of boys and girls is almost the same. Since all learners within the school were eligible to participate in the study, the class registers were used to randomly assign learners to the three experimental groups.

7.1. DATA ANALYSIS AND INTERPRETATION OF RESULTS

The hypothesis $H_0$: There will be no significant difference in the scores of participants required to write composition under different time limits.

The hypothesis $H_0$: $\mu_1 = \mu_2 = \mu_3$ was tested by a one way between groups ANOVA to find out if the means are significantly different and the results are presented on Table 1. Participant who wrote the composition under the 90 minute time limit obtained an average score of 6.46 or 65%, this average is the highest when compared to the other two means. The analysis of data indicated that there is a significant difference ($F(2) = 4.82, p < .05$) in the in the scores of learners in English composition various time limits. Therefore, the null hypothesis is rejected. It is important to note that the 30 minute time limit group performed better than the 60 minute one.

Table 1
One-way between-groups ANOVA of the Student Performance in English Composition Under Different Time Limits

<table>
<thead>
<tr>
<th>Levels of Test Duration</th>
<th>n</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>36</td>
<td>6.00</td>
<td>1.55</td>
<td>.26</td>
</tr>
<tr>
<td>60</td>
<td>26</td>
<td>5.27</td>
<td>1.34</td>
<td>.26</td>
</tr>
<tr>
<td>90</td>
<td>28</td>
<td>6.46</td>
<td>1.32</td>
<td>.25</td>
</tr>
<tr>
<td>Total</td>
<td>90</td>
<td>5.93</td>
<td>1.48</td>
<td>.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean squares</th>
<th>F</th>
<th>$p &lt;$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Groups</td>
<td>19.52</td>
<td>2</td>
<td>9.76</td>
<td>4.82</td>
<td>.010</td>
</tr>
<tr>
<td>Within Groups</td>
<td>176.08</td>
<td>87</td>
<td>2.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>195.60</td>
<td>89</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The probability that the observed differences in the means of the cognitive levels would have occurred by chance if the null hypothesis were true is less than .05. To find out where the differences lay, a post-hoc Least Significant Difference (LSD) test was performed in order to determine the levels or pairs of means that were significantly different.

As Table 2 show, the performance of learners at 90 minutes composition writing duration (mean difference = 1.20, $p < .05$) is significantly different from when is 60 minutes duration. The null hypothesis of no difference in the performance of participants required to write composition under different time limit conditions is therefore rejected. The results show that time limit has a significant effect on performance with the 90 minutes group significantly outperforming the 60 minutes group.
There will be no significant difference in the scores of participants required to write a letter under different time limits.

The hypothesis $H_0: \mu_1 = \mu_2 = \mu_3$ was tested by a one way between groups ANOVA to find out if the means are significantly different and the results are presented on Table 3. It is evident from Table 3 that the 90 minute group has maintained its superiority over the other groups by obtaining an average of 5.29 (53%). The analysis of data indicated that there is a significant difference (F (2) = F 5.96, $p < .05$) in the scores of participants required to write a letter under different time limits. The situation has markedly changed with respect to the 30 minutes and 60 minutes groups. Specifically, the 60 minutes group has outperformed the 30 minutes group by obtaining an average score of 4.5 (45%) while the 30 minutes group lacked behind with an average score of 3.5 (35%).

Table 3

<table>
<thead>
<tr>
<th>Levels of Test Duration</th>
<th>n</th>
<th>Mean</th>
<th>Std. Deviation</th>
<th>Std. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>36</td>
<td>3.50</td>
<td>2.21</td>
<td>.37</td>
</tr>
<tr>
<td>60</td>
<td>26</td>
<td>4.50</td>
<td>2.44</td>
<td>.48</td>
</tr>
<tr>
<td>90</td>
<td>28</td>
<td>5.29</td>
<td>1.41</td>
<td>.27</td>
</tr>
<tr>
<td>Total</td>
<td>90</td>
<td>4.34</td>
<td>2.18</td>
<td>.23</td>
</tr>
</tbody>
</table>

Source of Variation | Sum of Squares | df | Mean squares | F | $p <$
Between Groups | 51.11 | 2   | 25.55        | 5.96 | .004 |
Within Groups | 373.21 | 87  | 4.29         |     |       |
Total | 424.32 | 89  |              |     |       |

To find out where the differences lay, a post-hoc Least Significant Difference (LSD) test was performed in order to determine the levels or pairs of means that were significantly different. It is clearly evident that as the time limit is increased from 30 minutes to 90 minutes, there is a corresponding increase in the performance level of participants. As Table 4 show, the performance of learners at 30 minutes letter writing duration (mean difference = -1.79, $p < .05$) is significantly different from when is 90 minutes duration. Therefore, there is a significant difference in the performance of participants as a function of time. In other words, as the duration of writing the letter is increased, the marks of the participants also in increase.

Table 4

<table>
<thead>
<tr>
<th>(I) time</th>
<th>(J) time</th>
<th>Mean Difference (I-J)</th>
<th>Std. Error</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>60</td>
<td>-1.00</td>
<td>.53</td>
<td>.15</td>
</tr>
<tr>
<td>90</td>
<td>60</td>
<td>-1.79</td>
<td>.52</td>
<td>.00</td>
</tr>
<tr>
<td>30</td>
<td>90</td>
<td>1.00</td>
<td>.53</td>
<td>.15</td>
</tr>
<tr>
<td>90</td>
<td>90</td>
<td>.79</td>
<td>.56</td>
<td>.35</td>
</tr>
<tr>
<td>30</td>
<td>60</td>
<td>.79</td>
<td>.56</td>
<td>.35</td>
</tr>
</tbody>
</table>

$H_0$: The time limit effect will have no significant impact on the correlation between composition and letter writing scores.

$H_0: \rho_1 = 0$

This hypothesis was tested by correlating scores resulting from composition writing with letter writing. Resultant correlation values are not significant at .05 alpha level (see Table 5) showing that performance at 90 minutes duration for letter writing is significantly different from for composition writing. Table 5 shows that the correlation of composition and letter scores under the 90 minutes experimental condition has a value of .643 and this value is statistically significant.

Table 5

<table>
<thead>
<tr>
<th>composition</th>
<th>letter</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson</td>
<td>.64</td>
</tr>
<tr>
<td>Correlation</td>
<td>.00</td>
</tr>
<tr>
<td>Sig. (2-tailed)</td>
<td>28</td>
</tr>
</tbody>
</table>

**Correlation is significant at the 0.01 level (2-tailed).**
8. DISCUSSIONS, CONCLUSION AND RECOMMENDATIONS

It can be argued at this juncture that the 90 minutes allowed participants to demonstrate fully what they know, understand and can do. There is also ample evidence showing that the 30 minutes time limit put the participants under a lot of pressure. Most of the participants managed to use the 30 minutes available to write a full composition but there was no time left for them to write the letter. As a result, the 30 minute time limit learners did very well in the composition writing but their performance significantly dropped in the case of letter writing. The devotion of the entire 30 minutes time on composition writing may be a reason why the 30 minutes group obtained a higher mean score than the 60 minutes group. In agreement is Flaugher (1970) when he identified three potential sources of unfairness; these are the content sampled, the way in which the test is administered, and the way in which the scores are used. All of these three points pose a serious threat to test reliability validity. Specifically, unfairness due to the way in which the test is administered may largely be attributable to time limit issues. Enforcing time limit actually means that candidates are being assessed on a totally different construct as the imposition of severe time constraints induces an altogether different mode of writing called ‘assessment genre’. This renders questionable the comparability of the writing skills displayed in timed and untimed contexts” (Lau, 2013).

The analysis demonstrated that time limit had a pronounced effect on the relative standing as the value for 30 minutes and 60 minutes groups were very low. On the other hand, the correlation index for the 90 minutes group was substantial and significant. This not only demonstrates the superiority of the 90 minutes time frame but also attest to the high psychometric properties of the duration over the 30 minutes and the 60 minutes test administration procedures. According to Messick (1989), time limit imposed on an examination constraints examinees from demonstrating their language proficiency, then time limit can be considered an extraneous factor. Therefore, test scores generated under such conditions and the inferences derived from such test scores runs the risk of being unreliable and invalid. The evaluation of candidate’s work may not be a true reflection of his or her content mastery level. Under the 90 minutes time limit, candidates were able to demonstrate their writing proficiency in both composition and letter as indicated by the high correlation index between the scores.

9. Conclusion

Arbitrarily imposed time limits on examinations can have detrimental effects on the performance of candidates. Time pressure compromises the measurement process as the candidates are eventually assess and graded on how fast they can complete a given task. Students who are highly proficient in continuous writing but not necessarily fast are disadvantaged and credit is given to those who are able to write fast but at the same time less proficient in the art of writing. The candidates’ ability to write fast then becomes the attribute of interest rather than language proficiency. At the end of the day, construct irrelevant variance is introduced in the measurement process thus compromising the reliability and validity of the scores. Generating scores that have low reliability index further compromises educational reforms introduced in the early 1990s to improve the quality of basic education in Botswana. For criterion referenced testing to deliver the goods, candidates should be given enough time to put down what they know, understand and can do so that their strengths and weaknesses can be identified. The quality of education can only improve if the examination process generates reliable feedback information so that appropriate remedial and enrichment actions can be undertaken. Increasing the time limit for composition and letter writing to 80 or 90 minutes would go a long way in creating an environment under which candidates can demonstrate their full potential.

10. Recommendations

Botswana Examination Council should increase the time limit for letter and composition writing to 90 minutes to improve student performance since that will minimise extraneous factors contributing towards reaching their full potential.

11. References


