New Zealand Studies in Applied Linguistics, 2018, 24 (2), 24-41

GROUP VS. INDIVIDUAL: HOW ADMINISTRATION PROCEDURE CAN AFFECT VOCABULARY TEST SCORES

Averil Coxhead[1], I.S.P. Nation[2], Lisa Woods[3] & Dalice Sim[4]

Victoria University of Wellington[1, 2, 3]
Otago University[4]

Abstract

The primary aim of this study was to investigate two ways of administering a vocabulary size test (individual versus group administration) in New Zealand secondary schools. Two equivalent forms of the 20,000 version of the Vocabulary Size Test were used in this study. One hundred and eleven 13 to 17 year old native speakers of English at secondary school took one form of the test under group testing conditions. That is, each student took a form of the test whilst sitting in a room with other students who were also taking a form of the test. Each student also took an individually-administered form of the test. For a majority of the test-takers, scores on the individually-administered test were higher; in almost one-third of the cases substantially so. Factors affecting the test results were investigated, including the order of test administration, age, school year, and gender. The effort that a learner puts into taking a test can have a major effect on test scores and teachers need to take account of this when administering tests and interpreting test scores. Implications for pedagogy and options for further research based on this exploratory research are discussed.

Keywords: individual vs. group testing, test-taking effort, motivation, Vocabulary Size Test, secondary schools

Introduction

Pearson, Hiebert and Kamil (2007, p. 282) state that vocabulary assessment is 'grossly undernourished' and call for more practical and theoretical research in the field. Vocabulary size testing is no exception to this (Nation & Coxhead, 2014; Nation & Webb, 2011) and is beset with methodological errors (Nation, 2013). This article focuses on a practical issue in assessment, investigating whether administering a test individually or in groups has any impact on the results. This investigation was a core part of the methodology of a wider study of vocabulary size testing in New Zealand secondary schools in 2011 (see Coxhead, Nation & Sim, 2014; 2015) using the Vocabulary Size Test (VST) (Nation & Beglar, 2007). Copies of the test are available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx. In this study, participants took two different forms of the 20,000 version of the VST under two different testing conditions: group vs individual. To the best of our knowledge, this

practical issue has not been addressed in vocabulary testing research in the literature to date.

Individual administration versus group-test administration

For obvious time and resource-saving reasons, most tests are typically administered to a large group of learners (a whole class) at the same time with one administrator. Unfortunately, the scores on a group-administered test may depend not only on what the learners know, but also on how much effort they put into sitting the test. Informal work reported in Nation (2007) with the Vocabulary Levels Test (Nation, 2013; Schmitt, Schmitt & Clapham, 2001) in New Zealand schools made us aware that many learners did not try their best when sitting low stakes tests. When the whole class sat a test administered by the teacher, the learners did not try their hardest and their attention was not strongly focused on the test. As a result, their test scores did not fully reflect their knowledge. The learners tended to rush through the test, not using sensible test-taking strategies, and generally not giving it their best shot. The learners who were most likely to do this were those who tended to be struggling at school, and their poor results on the poorly taken test reinforced both their and their teachers' impressions of their inadequacy.

When a test administrator sat with an individual learner as the learner sat the test keeping them on task, their scores were much higher (Nation, 2007). It was thus clear that for some learners the results of a group-administered test were at best misleading and at worst strongly demotivating. Because we wanted to get the clearest picture of learners' vocabulary size, we needed to make sure that our test gave learners the best chance of showing what they really knew. Although individual test administration is extremely time-consuming, there was no point in giving group-administered tests if the results were meaningless for many of our test takers. Thus, a very important aim of our research was to compare the effects of group administered and individually-administered tests.

Test-taking effort

It is self-evident that the more attention learners give to a task, the better they will do it. While there is no research on this in vocabulary testing, there is evidence supporting it in assessment in a variety of disciplines. Wise and DeMars (2005) in a review of studies define *test-taking effort* as “a student’s engagement and expenditure of energy toward the goal of attaining the highest possible score on the test” (p.2). The term *test-taking motivation* is used in some studies. Particularly on low-stakes tests, a student’s score depends not only on the knowledge they bring to the test, but also on their test-taking effort (Barry, Horst, Finney, Brown & Kopp, 2010; Segal, 2012; Wise & DeMars, 2005, 2010). This effort can be largely explained by an expectancy-value model where the motivation with which a student approaches a task

depends on their expectancy of being able to complete the task successfully and the value that they see in completing the task (Wigfield & Eccles, 2000). If students see the task as being one they are capable of doing, and if they see some intrinsic or extrinsic value in doing the task well, they are likely to do the task well. If, however, they see the task as being too difficult and/or not being of any interest or value to them, they are likely to skip through the task as quickly as possible not giving it their full attention. Low motivation leads to low test scores. Wise and DeMars (2005) point out that test-taking effort also affects test validity. If students do not truly show what they know, then the test is not fully measuring what it is supposed to measure.

Researchers have investigated test-taking effort in a variety of ways. It has been measured using questionnaires (Barry et al., 2010; Liu, Bridgeman & Adler, 2012), and item response times on a computerised test (Wise & DeMars, 2010). Segal (2012) measured test-taking effort experimentally by getting students to do a matching task that did not involve previous learning but depended on how well they devoted themselves to the task. A variety of methods have been used to increase test-taking effort, such as providing monetary rewards (O'Neil, O'Neil, Abedi, Miyoshi & Mastergeorge, 2005), making assessment count in awarding grades or reporting results (increasing the stakes) (Liu, Rios & Borden, 2015), making tests easier, making tests interesting and challenging, and providing feedback. Another possibility is to individually monitor learners as they sit the test so that their performance is continually observed and commented on. This is the approach used in the present study.

In our study, we made efforts to maximise test-taking effort by making sure the test takers were focused on the test, paying attention, taking it seriously, keeping on task by getting help with problems that arose during the test and encouraged to keep working. Test takers working without this support may also remain engaged and on task, but there is a greater likelihood that this will happen if test takers are individually supported.

We hypothesize that for some learners, scores on a group-administered test are just as likely to reflect test-taking effort as knowledge. If this is true, we would expect that learners with low scores for their age and school year on a group-administered test would have significantly higher scores on an individually administered test.

The Vocabulary Size Test

The main focus of this article is not the Vocabulary Size Test, but two ways of administering the test. The VST was initially developed as a tool for measuring the written receptive vocabulary size of native speakers of English as well as speakers of English as a second or foreign language. Such measures can inform decisions around teaching and learning goals, the preparation of materials for learning and teaching, and curriculum design (Nation & Webb, 2011; Nguyen & Nation, 2011).

The VST uses a multiple-choice format where the target word is part of a short, non-defining sentence, as in the example below for the word *strap*.

strap: He broke the <strap>.
a promise
b top cover
c shallow dish for food
d strip of strong material

Figure 1 An item from the VST for the target word *strap*

The context helps with identifying the part of speech of the target word and limits the sense of the word. The vocabulary of the multiple-choice options and the sentences containing the target words is tightly controlled. Because of its multiple-choice format, the VST has come in for criticism because of the opportunities the format provides for uninformed guessing, and the uncertainty around the strength of knowledge needed to correctly answer a multiple-choice item (Gyllstad, Vilkaite & Schmitt, 2015; Stewart, 2014). Research has investigated different options for testing using the VST, including a study by Zhang (2013) investigating the effect of including an "I don't know" option for test-takers, and the development of bilingual forms of the VST (see Elgort, 2013; Nguyen & Nation, 2011; Elgort & Coxhead, 2016).

The original version of the VST contains 140 items, selected by a sampling from the first 14,000 frequency lists of Nation's (2006) British National Corpus (BNC) word family lists. The most frequent items are in the first 1,000 word family list, the next most frequent are in the second 1,000 list, and so on. The format of the VST might be a factor which affects test-taking effort. The items most likely to be known come first and the later items are much less likely to be known. This may result in learners giving up when they move into the more difficult parts of the test. Initial analysis of the computer-based form of the test on Myq Larson's online VST (www.my.vocabularysize.com), where high frequency items are spread through the whole test, shows test takers take the same amount of time on all of the questions in the test. They do not seem to rush through the more difficult parts of the test. The test takers for the online VST are generally not the same kind of learners who raised the concerns of the test developers originally. The current analysis focuses on the frequency-sequenced form of the test.

The VST was developed using corpus-based approaches and has been adapted for computer-based testing. The original version was evaluated by Beglar (2010). The test has been extended to sample from 20,000 word families and six forms of the 20,000 test have been developed (Coxhead, Nation & Sim, 2014). These new forms

of the test each contain 100 items, with a sampling rate of five per 1,000 word families from the first twenty levels of Nation's BNC lists, based on findings from Beglar (2010) which suggest that a smaller sample size does not affect the overall results of the test. The purpose of this expansion of the test was to enable testing of the vocabulary size of native-speakers of English (Coxhead, Nation & Sim, 2015) and high proficiency learners of English (Coxhead, Nation & Sim, 2014).

Coxhead, Nation and Sim (2014) compared the results of 46 test-takers (31 university students and 15 professionals or retirees) who took all six forms of the test individually with a researcher alongside. Results showed that the six forms fell roughly into two groups of three tests (A, B and D; C, E, and F). Two forms of the test that were shown to have statistically comparable means were used in the present study: forms C and E. Forms A and B of the VST were already in the public domain, whereas C and E were in-house versions.

The 20,000 version of the VST can take between 20 and 60 minutes for a participant to complete. Each test-taker took as long as they needed to complete the test. The mean scores and standard deviations of the two tests were comparable to each other: Form C, M = 78.74, SD = 15.221; Form E, M = 78.20, SD = 13.341. In terms of raw scores, the means differ by less than 0.6 of a percent (Coxhead, Nation & Sim, 2014).

What other factors might affect test performance?

The Coxhead, Nation and Sim (2014) study suggested that participant variables such as gender, age, education level, and first language of the test takers did not affect the equivalence of the tests. Of the 46 test takers in that study (a convenience sample of 34 females, 12 males), 28 were first language speakers of English and 18 were second or third language speakers of English; aged between 16 and late 60s). Unsurprisingly, vocabulary size increases with age (Biemiller & Slonim, 2001; Coxhead, Nation & Sim, 2015; Farkas & Beron, 2004). Gender has been a focus of research into vocabulary, with Scarcella and Zimmerman (1998) finding that male English as a Second Language (ESL) students scored higher on a test of academic lexicon. However, Biemiller and Slonim (2001) did not find any difference between males and females in their vocabulary size testing (p. 502), and Coxhead, Nation and Sim (2015) did not find any difference for gender in their VST study, based on the same dataset as this present study of group vs. individual test conditions.

Research questions

1. Does group administration of the Vocabulary Size Test result in lower scores than individual administration?
2. Does the order of test administration affect group-administration versus individual-administration results

3. Do gender, age, school year, or school decile affect group-administration versus individual-administration results?
4. Which students are most affected by a group-administered test?
5. What proportion of students is affected by the method of testing?

Methodology

This research took place during normal school hours in regular school classrooms, computer classrooms, or libraries, depending on what was available at the time. Both computer-based and paper-based forms of the test were used, depending on the availability of computer facilities and the timing of the testing at each school. Where possible, computer-based tests were used rather than paper-based tests.

Group and individual testing conditions and test order were assigned randomly. In the individual administration, an administrator sat next to the test-taker as she or he answered the test questions on a computer or on paper. The administrator gave encouragement and praise, pronounced words when asked, and generally kept the learner on task. In the group administration, the test-takers answered the test questions on computer or on paper in groups of between six and thirty people in classrooms. Note that group size in the group administration of the test was not explored as a variable in the present study. The administrator handed out the tests, monitored the participants, answered any questions about the test including pronunciation of words if requested (except for the meaning of the test items), and collected the scripts. In the individual administration each test-taker had an administrator sitting next to them who supported only that individual. The contrast in this study is between individual or group test administration as a proxy for level of support and encouragement in terms of engagement and effort in the testing task.

At the end of each test, test takers were given their final score, that is, their vocabulary size result. They were also given a feedback sheet with suggestions on ways to increase vocabulary size (see Appendix One for an example). Messick (1996) recommends that test results and their interpretation should be given to test takers to encourage positive washback. Test takers were also given time with researchers to ask questions about the interpretation of their results at the end of the test. A total of 103 test-takers took the tests on the same day, six took the second test one day later, and two took the second test two days later.

Participants

The participants in this study ranged in age from 13 to 17. They came from eight schools in Aotearoa/New Zealand in Palmerston North and Wellington. In New Zealand, schools are placed into deciles according to the income of the parents and caregivers in the school area. Those schools rated as Decile 1 have the largest number

of socially and economically disadvantaged students. These deciles determine the amount of government funding, with lower decile schools getting additional funding. The schools in the study ranged from deciles 6 to 10, with the majority in deciles 6 and 9, so they were average to high income schools. A total of 67 participants came from a decile 6 school, and 42 came from a decile 9 school. Over 700 students in total took part in a wider study of vocabulary size testing. Out of the 700, a total of 111 secondary school students who were native speakers of English sat two forms of the test under the two different administration conditions. The number of females in the study was 53 and the number of males was 58. At seven schools, volunteers were called for through teacher networks, while at one school, students were pre-selected by a teacher to ensure a range of participants across the entire population of students. Ethics permission was sought and obtained from parents for test takers under the age of 16 and from the test takers themselves for those over the age of 16. Participants were told that the purpose of the tests was to measure vocabulary size and that, where possible, they were asked to take a group and an individual test.

Research assistant training

Sitting with a test-taker is a time-consuming task. Ten research assistants were trained and took part in this research project with the lead researchers. The research group worked together in schools, which meant that there was plenty of support and resources for the testing. Instructions for the research assistants for the project are in Appendix Two.

Results and Discussion

Research question 1: Does group administration of the Vocabulary Size Test result in lower scores than individual administration?

Here we only look at those students whose age "matched" their school year, i.e., year 9 and aged 13 or 14; year 10 and aged 14 or 15; year 11 and aged 15 or 16; year 12 and aged 16 or 17; year 13 and aged 17 or 18. A total of 111 students met these criteria (see Table 1).

Table 1 Number of native-speakers sitting both the individual and group administered tests by age and school year

		School Year					Total
		9	10	11	12	13	
Age	13	19	0	0	0	0	19
	14	19	24	0	0	0	43
	15	0	19	9	0	0	28
	16	0	0	2	11	0	13
	17	0	0	0	4	4	8
Total		38	43	11	15	4	111

It is important to control for age and school year because both of these factors are likely to affect vocabulary size.
A paired-samples t-test was used to determine if there was a statistically significant mean difference between the individual and group scores. Data from the 111 participants was approximately normally distributed, assessed by visually examining a Q-Q plot and boxplot of the differences. Participants had a higher score when taking the test as an individual (M = 59.51, SD = 11.659) as opposed to taking the test in a group (M = 56.18, SD = 13.245), a statistically significant mean difference of 3.33 points, 95% CI [1.91, 4.76], $t(110) = 4.46$, $p < .001$, $d = 0.44$. A total of 3.33 points on the test is equivalent to 660 word families in vocabulary size (3.3 times 200). The answer to research question 1 is that overall there is a difference between group-administered and individually administered test scores, with most gaining higher scores on the individually administered tests. However, the order of the testing may be an important factor.

Research Question 2: Does the order of test administration affect group-administration versus individual-administration results?

We used an independent samples t-test to compare the administration order of the tests. There were 80 students that had their group test first (administration order) and 31 had their individual test first. Data was approximately normally distributed, assessed by visually examining Q-Q plots and boxplots, and there was homogeneity of variances, as assessed by Levene's test for equality of variances (p = .381). Students who had their group test first had a higher mean difference (M = 4.20, SD = 8.04) than those who had their individual test first (M = 1.10, SD = 5.67). This difference of 3.10 points, 95% CI [-0.03, 6.23], was not statistically significant, $t(109) = 1.964$, p = .052, but a medium-sized effect was observed (d = 0.42). Students who took the individually administered test first only scored 1.10 points higher in this condition than in the group condition. This is a test-administration-order effect of basically 3.10 points, or about 620 word families. There is thus evidence of an order effect on the tests with the second test gaining a higher than expected score, presumably as a result of practice with the format. It was unlikely that there would be a fatigue effect as the two tests were generally not given one immediately after the other. The words tested in the two tests were different but from the same frequency levels.

We had fewer test takers overall who had individual testing first and group second. It is not surprising that more people took the group tests first, with small numbers of researchers available to sit with participants individually at any time (compared with the number of participants who could take group tests) and whether participants who had taken a group test were able to commit to taking an individual test because of time constraints in the busy school curriculum.

We looked into the form of the test (C or E) completed in each administration of the test (group or individual) to determine if both forms of the test were used equally as often in each administration of the test, see Table 2. Pearson's chi-square test of independence (with Yate's continuity correction) determined that whether or not the tests were taken as a group or individually is not independent of the order the tests were taken; $X^2 = 5.294$, $df = 1$, $p = .021$. There is a moderate association between the test administration and test order (*Phi* = 0.239, $p = .012$). In the group-first-individual-second test condition, 65% students took test C as a group (and test E individually). In the individual-first-group-second condition, 61.3% of students group took test E as a group (and test C individually). In a future study it would be wise to have equal numbers of different test forms in the two orders.

Table 2 Contingency table for Chi-Square tests

		Test	
		Group = C, Individual = E	Group = E, Individual = C
Order	Group first, Individual second	52 (65.0%)	28 (35.0%)
	Individual first, Group second	12 (38.7%)	19 (61.3%)

Research Question 3: Do gender, age, school year, school, or school decile affect group-administration versus individual-administration results?

We next asked if the size of the difference between individual and group scores differed by any of the variables we measured for these students. For each of gender, age, school year, school, and school decile we used a one-way ANOVA to compare the mean difference (between individual and group results) between groups of each factor (see Table 3). Data was approximately normally distributed, assessed by visually examining Q-Q plots and boxplots, and there was homogeneity of variances, as assessed by Levene's test for equality of variances ($p > 0.05$ for all factors except for school, where Welch's ANOVA is reported instead).

The difference between a student's individual and group scores did not depend on their gender, age, school year, school, or school decile. That is to say, the improvement in score of individual over group testing was present across all these different variables.

Table 3 Summary table of results of variables for the difference between group and individual administration

Variable	Comparison	*n*	Mean Diff	SD.	*F*

Gender	Female	53	2.53	7.49	$F(1, 109) = 1.15$,
	Male	58	4.07	7.62	$p = .286$
Age	13	19	1.74	6.88	$F(4, 106) = 0.81$,
	14	43	3.67	6.70	$p = .524$
	15	28	3.00	8.60	
	16	13	6.23	10.03	
	17	8	1.75	4.89	
School year	9	38	3.71	7.20	$F(4, 106) = 0.30$,
	10	43	2.51	7.95	$p = .876$
	11	11	3.36	4.70	
	12	15	4.87	9.78	
	13	4	2.75	5.62	
School	1	2	8.50	2.12	$F(6, 13.30) = 1.56$,
	3	31	4.00	7.43	$p = .234$
	4	17	2.88	4.69	
	5	6	2.83	17.76	
	6	13	3.54	9.93	
	7	34	2.79	5.97	
	8	8	2.75	6.07	
School Decile	6	67	3.52	8.51	$F(6, 108) = 0.59$,
	8	2	8.50	2.12	$p = .554$
	9	42	2.79	5.92	

For the gender variable as shown in Table 3, the mean difference between the group and individual tests for females was 2.53 and for males was 4.07. Although the difference was bigger for males, it was not significantly different ($p = .286$). The table also shows that standard deviations are quite large, indicating that the differences between the individual and group tests varied a great deal from student to student. The mean difference (individual – group) was not different by gender, age, school year, school, and school decile. That is, none of these variables influenced the distinction between individually administered and group-administered tests.

Research Question 4: Which students are most affected by a group-administered test?

The next step was to see if the effects of the two ways of administering the test were different for those students who had relatively low scores on the group-administered Vocabulary Size Test for their age compared with those who had relatively high scores. To make this comparison, the students at each of five ages (13-17 years old) were divided into quartiles based on their scores on the group-administered test. We were interested in whether or not the average size of the difference between individual and group scores was different for weaker or stronger students. Since these tests are most often conducted in a group setting, and since "normal" ranges are most often quoted by age, we split the students into quartiles based on their group scores and their age. That is, students aged 13 who were in the lowest 25% of their age group were designated as quartile one, as were 14 year old students who were in the

lowest 25% of their age group, and so on for ages up to 17. Students aged 13 who were in the second 25% were placed in the second quartile as were students aged 14 who were in the second 25% of their age group and so on. So in each quartile there were students from all age levels. We then used a paired t-test to compare group and individual means within each age quartile (Table 4). For each age quartile group, data was approximately normally distributed, assessed by visually examining a Q-Q plot and boxplot of the differences

Table 4: Paired samples statistics for the four quartiles based on group test scores for each age level

Quartile	*n*	Individual mean (SD)	Group mean (SD)	Mean difference (95% CI)	Paired t-test
1	28	49.29 (9.038)	40.93 (10.583)	8.36 (4.40, 12.32)	$t(27) = 4.33$, $p < .001$, $d = 0.82$
2	27	54.48 (8.976)	51.70 (6.354)	2.78 (0.45, 5.11)	$t(26) = 2.45$, $p = .021$, $d = 0.47$
3	30	62.97 (8.373)	61.13 (5.367)	1.83 (0.01, 3.67)	$t(29) = 2.04$, $p = .051$, $d = 0.37$
4	26	71.77 (5.458)	71.54 (4.320)	0.23 (-2.11, 2.57)	$t(25) = 0.20$, $p = .841$, $d = 0.04$

Using the ages to define the quartiles (based on the group results), the individual test score is significantly higher for quartiles one and two, and there is a small-to-medium sized effect observed in the third group. For quartile one (those with the lowest group test scores for their age group) there was a difference of over 8 points which is a difference in vocabulary size of 1600 word families. This is a big difference, and the difference was statistically significant (paired t-test, $p < .001$ see Table 4). For the second quartile, the individual results were on average 2.78 points higher than the group scores, and this difference was statistically significant ($p = .021$). For the third quartile, the individual results were on average 1.84 points higher than the group scores, and this represents a small-to-medium sized effect ($p = .051$). For the highest quartile, the individual results were on average 0.23 points higher than the group scores but this difference was not statistically significant ($p = .841$). The fourth quartile students did not benefit from having individual testing (the effect size was close to 0; d=0.04).

Clearly the effect of individual administration is strong enough to make a big difference for the lowest 25% of the students, and this effect persists at least to a small degree for another 50% of the students. Converting points to word families means that 8.36 points is equivalent to 1672 word families, as noted above (8.36 times 200), 2.78 points to 556 word families, and 1.84 points to 368 word families. If we relate this difference to the group scores (40.93), then learners in the first quartile sitting the individually administered test (49.29) increased their scores by over 8%. If group-administered Vocabulary Size Tests are used, at least one-quarter of the

students in decile 6 to 10 schools like the ones in this study are likely to have misleading results greatly under-estimating their vocabulary size. These results are likely to be even more misleading in lower decile schools, because vocabulary size is closely related to socio-economic status (Farkas & Beron, 2004).

Research Question 5 What proportion of students is affected by the test administration?

The previous section showed how the lower quartile of students in particular was affected by the use of a group or individual test. However, quartiles are a rather coarse measure. Table 5 shows the range of differences for participants with higher individual than group scores on the VST, with a total of 67.5% (75 out of 111 participants) overall scoring higher on the individual test condition. Five participants (4.5% of the participants) got exactly the same scores on the two types of test administration.

Table 5 Number of students at a range of point differences with higher scores for the individual than the group administered tests

Point differences between tests	Number of students	Percentage of students (%)
1.00	9	8.1
2.00	6	5.4
3.00	10	9.0
4.00	8	7.2
5.00	5	4.5
5.5 to 10	23	20.7
> 10	14	12.6
Total	75	67.5

Note in column 2 of Table 5 that 42 (5+23+14) participants (37.8%) scored an extra 5 points or more on the individual condition over the group test condition. That means that for this particular group of students, a group-administered test would greatly under-estimate their vocabulary size. Table 6 shows that 31 students (29.9%) gained higher scores on the group-administered tests.

Table 6 Number of students at a range of point differences with higher scores for the group than for the individually administered tests

Point differences between tests	Number of students	Percentage of students
< -10	3	2.7
-10 to -5.5	6	5.4
-5.00	3	2.7

-4.00	4	3.6
-3.00	8	7.2
-2.00	4	3.6
-1.00	3	2.7
Total	31	27.9

As the previous section showed, these were generally the students with higher scores for their age and school year levels.

Implications for teachers

The scores of some students on the individually administered test are not the same as their scores on a group-administered test, possibly because their group scores are more strongly dependent on a lack of test-taking effort than knowledge. It is not easy to isolate the various effects of individual administration. However, experience in administering such tests suggests that the following ranked list of factors accounts for most effects. The most influential factor is given first. The first two factors cover test-taking effort.

1 The learners give their full attention to the task. This means that they do their best to answer the questions and are not distracted by other factors. This is a direct result of having the full attention of the test administrator.
2 The learners are encouraged to remain positive about their chances of answering items correctly. This is encouraged by positive comments from the test administrator.
3 The learners are able to clarify aspects of the task such as whether they should answer items they are unsure about and clarify aspects of individual items such as the pronunciation of some of the words. The test administrator can provide needed information excluding, of course, the meaning of the word.
4 The learners are encouraged to take a strategic approach to answering the test items applying what test-wiseness skills they have.

Individually administered tests are largely impractical for more than a few students. Teachers who are concerned about their low scoring students could possibly use a group administration first for screening purposes and then an individual administration to see more accurately what these students know and can do. They could also try some of the techniques suggested by O'Neil et al. (2005) and Liu et al. (2015) to increase students' test-taking efforts.

Limitations

There are several limitations in this study. An obvious limitation is the variation between computer-based and paper-based forms of the test, even though it was the same test. Another is the possible effect of the kinds of support offered to the test

takers. For example, what effect might offering the pronunciation of target words have on the overall test scores? Note that both individual and group testers were able to ask for support in this way. The data from this study came from mid-to-high decile schools. It is possible that the effect of test condition on test takers could affect more participants and to a larger extent in lower decile schools. It is important to note that at times it was not possible to fully separate the individual and group test takers on busy school premises. Researchers and individual test takers tried as much as possible to be distanced from the group test takers, but often we shared a classroom. Finally, it could be other factors that are related to the test itself which could have affected the test-takers' efforts, such as its format, the inclusion of low frequency lexical items, and the lack of connection to the school curriculum (thanks to an anonymous reviewer for that suggestion).

Future research

A useful avenue for replication would be setting up age groupings so that there is a clear gap between age groups. That could mean choosing only test-takers who were born in the first six months of the year so that the various age groups do not contain learners who were only a few days or a few weeks away from another age group classification. Perhaps further research could investigate individual vs group test administration in different contexts or subject areas (apart from vocabulary size research) and with a range of older or younger learners, and therefore help with the interpretation of these initial results. Follow-up interviews with test takers might also shed light on how they reacted to the testing in different conditions.

Another useful initial piece of future research would be an approximate replication study (see Porte, 2012, for more on replication studies in applied linguistics) with learners of the same age taking the tests. An advantage of this replication is that we could increase the sample size. A replication study could measure the time taken for each test administration, which would allow for an analysis with time as a factor. Exploring group size as a variable would be a logical extension of the group versus individual distinction, possibly looking to see whether there is a linear effect or a critical threshold, as an anonymous reviewer suggested. A replication study with adults would also allow us to reflect on the interpretations of the original study. It would be useful to carry out a conceptual replication using the randomised online order form of the VST. Future research could also include observation of test takers and research during test administration to find out more about the nature and amount of support given in the individual and group testing. Finally, triangulating test findings with motivation and/or test-taking strategy surveys might provide a way of substantiating the ranked list of factors in the implications section above more objectively, as an anonymous reviewer suggested.

Conclusion

For many of the learners in this study, the effects of individual administration are particularly striking. As Table 6 shows, 14 learners (12.6%) had much higher scores in individual administration (10 points out of 100 or higher). It is clear that for a substantial number of learners, the results on a group-administered test could greatly misrepresent their knowledge. There is no reason to think that this factor is peculiar to vocabulary size tests; it could apply to any group-administered test.

Acknowledgements

The authors wish to thank the schools and students who took part in this study. Taking two forms of the VST took some participants took up to two hours of their time, and we are very grateful for that effort. A hearty thanks also to our highly-skilled research assistants for their time and dedication to the project. This research was supported by a Victoria University Research Fund grant.

References

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342-363.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing, 27*(1), 101-118.

Coxhead, A., Nation, P. & Sim, D. (2014). Creating and trialling six forms of the Vocabulary Size Test. *TESOLANZ Journal, 22*, 13-26.

Biemiller, A., & Slonim, N. (2001). Estimating root word vocabulary growth in normative and advantaged populations: Evidence for a common sequence of vocabulary acquisition. *Journal of Educational Psychology, 93*(3), 498-520.

Coxhead, A., Nation, P. & Sim, D. (2015). Vocabulary size and native speaker secondary school students. *New Zealand Journal of Educational Studies, 50*(1), 121–135.

Elgort, I. (2013). Effects of L1 definitions and cognate status of test items on the Vocabulary Size Test. *Language Testing, 30*, 253-272.

Elgort, I., & Coxhead, A. (2016). An introduction to the Vocabulary Size Test: Description, application and evaluation. In J. Fox & V. Aryadoust (Eds.), *Trends in language assessment research and practice* (pp. 286-301). Newcastle upon Tyne, England: Cambridge Scholars Publishing.

Farkas, G., & Beron, K. (2004). The detailed age trajectory of oral vocabulary knowledge: Differences by class and race. *Social Science Research, 33*, 464-497.

Gyllstad, H., Vilkaite, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *International Journal of Applied Linguistics, 166*(2), 276-303.

Liu, O. L., Bridgeman, B., & Adler, R. M. (2012). Measuring learning outcomes in higher education: Motivation matters. *Educational Researcher, 41*(9), 352-362.

Liu, O. L., Rios, J. A., & Borden, V. (2015). The effects of motivational instruction on college students' performance on low-stakes assessment. *Educational Assessment, 20*(1), 79-94.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*, 241-256.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*(1), 59-82.

Nation, I.S.P. (2007) Fundamental issues in modelling and assessing vocabulary knowledge. In H. Daller, J. Milton & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 35-43). Cambridge: Cambridge University Press.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Nation, P. & Coxhead, A. (2014). Vocabulary size research at Victoria University of Wellington, New Zealand. *Language Teaching, 47*(3), 398-403.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.

Nguyen, L. T. C., & Nation, I. S. P. (2011). A bilingual vocabulary size test of English for Vietnamese learners. *RELC Journal, 42*(1), 86-99.

O'Neil, A. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*(3), 185-208.

Pearson, P., Hiebert, E., & Kamil, M. (2007). Vocabulary assessment: What we know and what we need to learn, *Reading Research Quarterly, 42*(2), 282-296.

Porte, G. (2012). *Replication research in applied linguistics*. Cambridge: Cambridge University Press.

Scarcella, R. & Zimmerman, C. (1998). Academic words and gender: ESL student performance on a test of academic lexicon. *Studies in Second Language Acquisition, 20*(1), 27-49.

Schmitt, N., Schmitt, D. & Clapham, C. (2001) Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing 18*(1), 55-88.

Segal, C. (2012). Working when no one is watching: Motivation, test scores, and economic success. *Management Science, 58*(8), 1438-1457.

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST? *Language Assessment Quarterly, 11*(3), 271-282.

Wigfield, A., & Eccles, J. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*, 68-81.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1-17.

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*(1), 27-41.

Zhang, X. (2013). The *I Don't Know* Option in the Vocabulary Size Test. *TESOL Quarterly, 47*(4), 790-811.

Appendix One
Vocabulary Size Test feedback sheet for scores over 10,000 words

Victoria
UNIVERSITY OF WELLINGTON
Te Whare Wānanga
o te Ūpoko o te Ika a Māui

Vocabulary Size Test Feedback Over 10,000 words

Thanks for taking this vocabulary size test. This is a test of your receptive vocabulary, meaning the knowledge of words you have for reading and listening. Generally speaking, if we count from when a native speaker is two or three years old, their vocabulary size increases by 1,000 words per year. For second language learners, this rate is not so fast. Your vocabulary size is over 10,000 words, which indicates that it is more than enough for university studies. If you want to increase it, you could try these things:

1. Increase the amount of reading you do every day.
2. Read widely. For example, you could read newspapers online from NZ and other countries, non-fiction books on topics that interest you, and magazines such as the Listener or the Economist, and your textbooks too.
3. Listen widely and often. For example, you could listen to Radio New Zealand National programmes on radio or as podcasts, watch information rich programmes on tv like documentaries, and talk about the content and ideas of these programmes with friends and family.
4. Concentrate on the technical vocabulary of your subjects.
5. Use activities such as word cards to keep track of words you want to learn and revise these words often. Pay attention to the spelling of words, as well as common patterns they occur in.
6. Try to introduce words you recognise but don't use often into your speaking and writing. Seek feedback on your use of these words from other people.
7. Look for examples of words being used in different contexts. Note them down (on word cards or in a notebook, for example) and focus on learning their meaning.

Appendix Two

Instructions for individual test administration for research assistants

Procedure

1 Sit next to the person doing the test so that you can see the screen. ONLY TEST ONE PERSON AT A TIME. This is because it is essential that the person sitting the test gives it their full attention, is personally encouraged to keep trying, and is helped if the pronunciation of any words is needed.

2 Provide support and encouragement in any or all of the following ways that you see as being appropriate. Pilot testing has revealed that a learner who sits the test with someone like you sitting next to them and encouraging them gets much higher scores than someone sitting the test without the support.

(i) Say "Good" after each item is done. If the learner wants feedback on correctness, provide the answer after they have completed the item.

(ii) Encourage the learner to rest after doing several levels of items. You could use this time to get feedback on how they view the test, and to confirm and check items on the background form.

(iii) If the learner is uncertain about an item, encourage them to guess. For many items, learners will have only a vague idea of their meaning and this knowledge may even be subconscious. Where possible encourage informed strategic guessing, but discourage what is obviously random guessing. Overall, however, it is better to guess than not guess.

(iv) If you think a learner has problems reading a word, read it aloud for them. The vocabulary size test is a test of receptive knowledge of vocabulary, but this does not have to be limited to reading.

(v) If you think that the learner is not sitting the test properly, or if you have any reason to doubt the accuracy of their performance, let us know. In some cases you may decide to abandon the administration to a particular learner if you feel the result will not truly reflect their knowledge. Be sure to keep some notes about this if you do.

(vi) It is important that the learner tries all items on the test if this is at all possible. Typically learners know some very low-frequency words.

3 Please note any comments you have about the test, about the learner, or about the learners sitting of the test. We want the results to reflect the learners' knowledge as closely as possible.