Wojciech Malec
Zamość - Lublin
malec.wojciech@pwszzamosc.pl

Language testing: towards a methodology for the study of the method effect

Sprawdzanie umiejętności w nauce języka: w stronę metodologii badania efektu metody

Streszczenie:

Artykuł podejmuje kwestię metodologii badania efektu metody w kontekście testów sprawdzających. Podczas gdy w tradycyjnych badaniach efektu metody w pomiarze różnicującym stosowana jest oparta na korelacjach macierz wielu cech – wielu metod, w przypadku pomiaru sprawdzającego kluczową rolę odgrywa istotność różnicy między średnimi, ponieważ trudność zadań testowych ma bezpośredni wpływ na decyzje klasyfikacyjne. Autor zwraca szczególną uwagę na strukturę eksperymentu oraz sposoby kontrolowania wpływu zmiennych zakłócających w planach z powtarzanym pomiarem.

Słowa kluczowe: testowanie języka obcego, efekt metody, pomiar sprawdzający, plany z powtarzanym pomiarem

Summary:

This article discusses the methodology for investigating the effect of item format on test performance. While correlational procedures have traditionally been used in examinations of the method effect in the context of norm-referenced tests, the significance of differences between means is more important for criterion-referenced testing because test difficulty has a direct influence on classification decisions. Special attention is given to ways of controlling confounding factors in repeated-measures experimental designs.

Keywords: language testing, method effect, criterion-referenced measurement, repeated-measures designs

1. Introduction

The last two decades have witnessed a burgeoning of scholarly interest in language testing, reflected in an enormous amount of research into the design and development of useful language tests. Numerous theoretical and empirical investigations have for the most part centred upon improving the quality of measurement instruments, particularly their reliability and validity. However, as noted by Bachman, "[w]hile this research has deepened our understanding of the factors and processes that affect performance on language tests, it has also revealed lacunae in our knowledge and pointed to new areas of research".[1]

The influence of test method on test performance is an acknowledged but still relatively under-researched problem. Although there is now substantial evidence[2] that the characteristics of the tasks themselves can make a significant difference to test scores, there is still considerable uncertainty as to what aspects of language ability are measured by the many task types used in language assessment. Moreover, it is not at all clear which testing techniques can be regarded as cognitively more demanding than others. Alderson *et al*. went as far as to suggest that a thorough understanding of the method effect is "the Holy Grail of language testing".[3] Pursuing it is undoubtedly well worth the effort because unless we learn to more fully understand the impact that a given method is likely to have on test performance, we cannot really develop useful language tests and appropriately interpret their results.

2. The effect of item format

The reason why we construct and administer language tests is that we want to assess language ability (or language knowledge). Inferences about the ability being measured are made on the basis of the test takers' scores. However, test score variance is never solely and exclusively due to variations in ability. A variety of confounding factors, both external (e.g. weather) and internal (e.g. motivation) can affect test

1 L. F. Bachman, *Modern language testing at the turn of the century: assuring that what we count counts,* "Language Testing", 2000 No. 1, p. 2.

2 E.g. J. C. Alderson and A. H. Urquhart, *The effect of students' academic discipline on their performance on ESP reading tests,* "Language Testing", 1985 No. 2; P. Arnaud, *Vocabulary and grammar: a multitrait-multimethod investigation,* "AILA Review", 1989; R. Freedle and I. Kostin, *The prediction of TOEFL reading comprehension item difficulty for expository prose passages for three item types: main idea, inference, and supporting idea items,* Research Report No. 93-13. Princeton, NJ 1993; M. Kobayashi, *Method effects on reading comprehension test performance: text organization and response format,* "Language Testing", 2002 No. 2; E. Shohamy, *Does the testing method make a difference? The case of reading comprehension,* "Language Testing", 1984 No. 2.

3 J. C. Alderson, C. Clapham and D. Wall, *Language test construction and evaluation,* Cambridge 1995, p. 45.

performance. One of the sources of variance that is not associated with language ability is the method of testing.[4]

Test method is a general term used to refer to the testing procedure as a whole and as such can be looked at and examined in its entirety. However, within the framework of test method facets, or task characteristics,[5] various aspects of the testing procedure can be delineated and analyzed separately – a researcher can focus primarily on only one of the test method facets. This might be, for example, the format of the test items and the way it impacts on the difficulty of the test, which is particularly interesting from the perspective of criterion-referenced (CR) classroom testing because item/test difficulty has a direct influence on the mastery/non-mastery decisions made on the basis of students' test performance. The present article discusses the methodology for investigating precisely this problem: the effect of item format on test performance in the context of criterion-referenced progress tests.[6] The hypothetical test construct which appears in all of the examples discussed in the following sections is knowledge of collocations, but the methodology is applicable to other constructs too.

3. Experimental design considerations

Before the construction of the measurement instruments, important decisions have to be made concerning the design of the experiment, whose purpose is to explore the effect of several item formats on the test takers' performance, as represented by test scores. On the face of it, the task seems fairly straightforward and uncomplicated, and requires an investigation of the relation between two variables: item format as the independent variable and test performance as the dependent variable.

However, the fact that the tests are meant to be CR classroom progress/achievement tests necessitates administering them to a group of students who are being taught the same syllabus, thereby limiting the possibility of obtaining a large sample. Consequently, the option of an independent-groups design is barely viable: several reasonably-sized groups would be needed, one for every item format. In general, 10 subjects for each group can be seen as the bare minimum, but the larger

4 Cf. L. F. Bachman, *Statistical analyses for language assessment,* Cambridge 2004, p. 156.

5 L. F. Bachman, *Fundamental considerations in language testing,* Oxford 1990; L. F. Bachman and A. S. Palmer, *Language testing in practice: designing and developing useful language tests,* Oxford 1996.

6 See also W. Malec, *The impact of item format on test performance in criterion-referenced assessment of collocations,* PhD thesis, KUL 2006; W. Malec, *Efekt metody w pomiarze sprawdzającym na przykładzie testowania kolokacji języka angielskiego,* in: *Uczenie się i egzamin w oczach uczniów,* B. Niemierko and M. K. Szmigel (eds.), Kraków 2007.

the sample, the higher the precision of the analysis and the greater the credibility of the conclusions.[7] Therefore, an alternative solution, a repeated-measures design, is preferred.

Besides requiring fewer participants,[8] the most important advantage of a repeated-measures design is that individual differences between participants are controlled, which results in a reduction of unsystematic variability and, by the same token, in greater statistical power to detect an effect. In other words, the fact that the same subjects are tested in several different experimental conditions increases the correlation between measurements and reduces the error term (i.e. the variance that is due to individual differences between the subjects rather than due to the experimental modification). Field demonstrated that the method of data collection "can make the difference between detecting an effect and not detecting one".[9] When he analyzed a set of data using the dependent *t*-test, the difference between means was statistically significant, whereas when the same data were analyzed using the independent *t*-test, no significant difference was found. What the foregoing amounts to is that, given a medium-sized sample of students, the great merit of a repeated-measures design, compared to an independent design, is a lower probability of making a Type II error (i.e. failing to detect an effect that does genuinely exist).

On the other hand, a serious drawback to repeated-measures designs is that they create the possibility of transfer between treatment conditions, which may invalidate the results of the experiment. More specifically, the effect of the experimental modification as observed for the second level of the treatment variable may not be independent of the first level "for reasons of fatigue, practice, or some other cause".[10]

To conclude, a repeated-measures design is a better choice than an independent design as long as measures are taken to eliminate any carryover effects which may otherwise nullify the impact of the experimental modification on the outcome (the dependent variable). Such measures usually involve randomizing the order of treatments for each participant or group of participants.

---

7 A. Radźko, *Skuteczność metod statystycznych i warunki ich stosowania w badaniach pedagogicznych,* in: *Zasady badań pedagogicznych: strategie ilościowe i jakościowe,* Rozdział 6, T. Pilch and T. Bauman, Warszawa 2001, p. 131.

8 G. A. Ferguson and Y. Takane, *Analiza statystyczna w psychologii i pedagogice,* Warszawa 1999, p. 371; P. Francuz and R. Mackiewicz, *Liczby nie wiedzą, skąd pochodzą: przewodnik po metodologii i statystyce nie tylko dla psychologów,* Lublin 2005, p. 64.

9 A. Field, *Discovering statistics using SPSS,* London 2005, p. 304.

10 G. A. Ferguson, *Statistical analysis in psychology and education,* New York 1971, pp. 201-202.

3.1 Refining the design

The choice of a repeated-measures design calls for consideration of the following further options:

A) A single-item-format test (e.g. fill the gaps) covering a set of collocations can be administered to a group of students. At a later time, the same collocations can be given to the same group of students in another item format (e.g. multiple choice), and so on until all of the item formats have been administered. The number of tests corresponds to the number of response formats.

B) The same as A, but to eliminate any systematic effects of order, for every student who is to take the tests, the order of the formats can be randomized.

C) The same as A, but instead of using the same set of collocations every time a new item format is administered, a different set of collocations can be used.

D) On a single test administration, the students can get several different sets of collocations, each set in a different item format.

E) All of the item formats can be administered in a counterbalanced manner such that no student gets the same collocation in any two different item formats.

Brown and Hudson pointed out that "frequently examinees remember items from the previous test, and that memory has a contaminating effect on any subsequent encounter with those items".[11] With this observation in view, there is little to support option A. In all likelihood, whichever item format is administered last will produce higher scores than the one administered first.[12]

Therefore, randomizing the order of the response formats, as stated in option B, seems an essential improvement. This method of avoiding the effects of order is very common in psychological and educational research.[13] In the experiment under discussion, thanks to randomization, no single item format would be privileged by being administered in its entirety at a later time than any other format. However, as noted by Brzeziński,[14] methodologically elegant though it may seem, randomizing

---

11 J. D. Brown and T. Hudson, *Criterion-referenced language testing,* Cambridge 2002, p. 128.

12 Ironically, if the time span between test administrations is very long, not only may the students forget items from the previous test, but they might also forget the meaning and use of some of the collocations that they studied before the first administration. In this rather unlikely event, a test format that is administered later can actually give rise to scores significantly lower than those derived from a test format administered earlier (cf. Bachman, 1990, p. 182). However, what is important for our purposes is that the students' knowledge of the collocations being tested is unlikely to be the same on any two administrations of the test, from which it follows that the order of treatments is almost certain to have some impact on test scores that is impossible to control in this design.

13 See, e.g., T. W. Pavkov and K. A. Pierce, *Do biegu, gotowi – start! Wprowadzenie do SPSS dla Windows,* Gdańsk 2005, p. 65.

14 J. Brzeziński, *Badania eksperymentalne w psychologii i pedagogice,* Warszawa 2000, p. 156.

the treatments is far from convenient because it requires increasing the sample size. Specifically, the number of different ways in which *n* treatments can be arranged in order equals *n*! (*n* factorial). Thus, with six item formats, there are 6! = 720 possible arrangements, also expressed as the number of permutations, of these formats.

Alternatively, the order of the formats can be the same for the whole group of students, but every time the test is administered, a new set of collocations can be used for assessment (option C). In this design, effects of order would clearly not be an issue. However, test format would be confounded with test content (the collocations): higher scores produced by, for example, multiple-choice items could not be interpreted to mean that the MC format is easier than the other formats. The different levels of difficulty of the formats as indicated by mean scores might actually be due to the fact that the collocations selected for one particular format were easier than those selected for the other formats. Again, the solution lies in randomization. Prior to administering the tests, the requisite number of collocations might be selected and then randomly assigned to several sets, thereby eliminating any systematic bias in mean scores that might be due to the collocations' varying degrees of difficulty. Nevertheless, option C is not free from the confounding effects of external and internal factors such as temperature, noise, fatigue, mood, motivation, changes in students' knowledge and ability from one test to the next, etc.[15] To put it differently, within-subjects variations in performance may not be due to the experimental modification only.

Option D addresses this problem: all the response formats are administered at the same point in time, i.e. in the same testing situation. In this design, the order of the formats does not really matter (students often do not follow the order anyway). Each format covers a different set of collocations, balanced in terms of difficulty thanks to random assignment. For the purposes of studying the method effect, this design is quite satisfactory, yet there is still one problem that needs to be addressed. Specifically, even small differences in difficulty between the sets of collocations may lead to invalid conclusions.

Randomization can be seen as "analogous to insurance, in that it is a precaution against disturbances which may or may not occur and that may or not be serious if they do occur".[16] In option D, randomization (here, random assignment of previously selected collocations to several sets) is supposed to minimize the biasing influence of the collocations' unequal difficulties. However, even though this influence can be minimized, it is unlikely to be eliminated altogether. By way of illustration, suppose

15 Cf. Brown and Hudson, p. 150.

16 W. G. Cochran and G. M. Cox, *Experimental designs,* New York 1957, quoted in Ferguson, p. 201.

we need to assign 30 collocations to three balanced sets. Suppose also that half of the collocations, notated a_1, represent one level of difficulty, and the other half, notated a_2, represent another level of difficulty. The task is to assign 15 a_1s and 15 a_2s to three sets of equal difficulty, i.e. sets containing 5 a_1s and 5 a_2s each. Using the RAND() function in Excel,[17] it can be demonstrated that the three sets are actually more likely than not to be unequal.

The exact probability of obtaining three equal sets through randomization can be found with the help of the following formula for permutations with repetitions[18]:

$$(1) \qquad P_n^{n_1,n_2,\ldots,n_k} = \frac{n!}{n_1!n_2!\ldots n_k!}$$

where: $P_n^{n_1,n_2,\ldots,n_k}$ = permutations of n elements in which k distinguishable elements are repeated in such a way that:
– element a_1 is repeated n_1 times (there are n_1 indistinguishable a_1 elements)
– element a_2 is repeated n_2 times (there are n_2 indistinguishable a_2 elements)
– element a_k is repeated n_k times (there are n_k indistinguishable a_k elements)

First, the number of possible arrangements of 5 a_1s and 5 a_2s in each set can be calculated as follows:

$$(2) \qquad P_{10}^{5,5} = \frac{10!}{5!5!} = \frac{3628800}{120 \times 120} = \frac{3628800}{14400} = 252$$

Second, the total number of possible arrangements of a_1 collocations and a_2 collocations in all three sets such that each set contains exactly 5 a_1s and 5 a_2s is the product of $P_n^{n_1,n_2,\ldots,n_k}$ for each set, or simply:

$$(3) \qquad \left(P_n^{n_1,n_2,\ldots,n_k}\right)^3 = \left(P_{10}^{5,5}\right)^3 = 252^3 = 16003008$$

Third, the total number of any arrangements of a_1s and a_2s in all three sets taken as a whole can be found by calculating the permutations of $n = 30$ elements in which $k = 2$ distinguishable elements are both repeated $n_1 = n_2 = 15$ times:

---

17 Microsoft Excel, Version 11.5612.5606, Microsoft Corporation 2003.

18 T. Koshy, *Discrete mathematics with applications,* Burlington, MA 2004, p. 428; H. B. Fine, *College algebra,* Providence, RI 2005, p. 398; K. G. Calkins, *Permutations with repeated elements.* [Retrieved July 5, 2006 from http://www.andrews.edu/~calkins/math/webtexts/prod02.htm#RPERM]

$$(4) \qquad P_{30}^{15,15} = \frac{30!}{15!15!} = \frac{2.65253\mathrm{E}+32}{(1.30767\mathrm{E}+12)\times(1.30767\mathrm{E}+12)} = \frac{2.65253\mathrm{E}+32}{1.71001\mathrm{E}+24} = 1.55\mathrm{E}+08$$

Finally, the probability that the three sets of collocations are equal as a result of randomization can be estimated. Within the framework of classical probability, the probability that Event A will occur (P[A]) is calculated in the following way[19]:

$$(5) \qquad \mathrm{P[A]} = \frac{\text{Number of possible outcomes in which Event A occurs}}{\text{Total number of possible outcomes}}$$

In our example, Event A occurs when there are 5 a_1 collocations and 5 a_2 collocations in each of the three sets. The number of possible outcomes in which Event A occurs has been calculated in (3), and the total number of possible outcomes has been found in (4). The probability that the sets are equal is as follows:

$$(6) \qquad \mathrm{P[A]} = \frac{16003008}{1.55\mathrm{E}+08} = .10317$$

Thus, we can be 10% confident that random assignment will result in equal sets of collocations.

Naturally, in a real situation, the collocations would most likely represent more than two levels of difficulty, and as the number of collocations which are distinguishable in terms of difficulty rises, the probability of obtaining unequal sets increases exponentially. For example, if our array of thirty collocations consisted of $6a_1+6a_2+6a_3+6a_4+6a_5$, i.e. five groups characterized by distinguishable levels of difficulty, each group containing six collocations of the same difficulty, the probability of successfully assigning them to three equal sets, each containing $2a_1+\ldots+2a_5$, would be P[A] = . $\left(P_{10}^{2,2,2,2,2}\right)^3 \div P_{30}^{6,6,6,6,6} = .00106$ Therefore, we could be 99.9% certain that the sets would be unequal as a result of randomization.

19 R. A. Donnelly, *The complete idiot's guide to statistics,* Indianapolis 2004, p. 76.

The above discussion is not supposed to undermine the value of randomization. In fairness, there is no better way of controlling bias in an experiment than by randomizing entities whose characteristics are unknown[20] (in our case, when the difficulty of the collocations is unknown). The purpose of the above calculations is simply to demonstrate how difficult it may be to obtain ideal equivalence of several sets of collocations.

The implications of having unequal sets of collocations are now discussed. Suppose that a group of 15 students take a test consisting of three test methods (M1, M2, M3) of equal difficulty, such that each method covers a different set of 10 collocations (see Table 1). Suppose further that the sets of collocations represent three levels of difficulty, such that SET II (marked light grey in the table) yields scores that are consistently lower (-1 point) than those produced by SET I (white), and that SET III (dark grey) yields scores that are consistently two points lower than those produced by SET I. Using a repeated-measures design, we can test the null hypothesis that the three methods are equivalent in terms of difficulty. If the variant of the design defined here as option D were used in this experiment (Table 1), then we would reject the null hypothesis, when, in reality, it is true (a Type I error). The weakness of this design is that it leads to a misleading result: the apparently significant differences between the means are actually due to the differences in difficulty between the sets of collocations, and not due to any differences between the methods.

20 However, for a discussion of the limitations of randomization in the social sciences, see S. Ackroyd and J. A. Hughes, *Data collection in context,* Harlow 1992.

Table 1. *Hypothetical test scores (option D)*

Student #	M1	M2	M3
1	8	7	6
2	9	8	7
3	10	9	8
4	8	7	6
5	7	6	5
6	8	7	6
7	9	8	7
8	10	9	8
9	8	7	6
10	7	6	5
11	8	7	6
12	9	8	7
13	10	9	8
14	8	7	6
15	7	6	5
Mean	8.4	7.4	6.4

Note: SET I (0); SET II (-1); SET III (-2)

There is a way of eliminating the small differences that are likely to exist between sets of collocations. Table 2 illustrates how the collocation test can be administered in a counterbalanced manner (option E). In this design, all the students are divided into as many groups as there are test methods, and sets of collocations. Each student in Group A gets ten collocations (SET I) in Method 1; another ten collocations (SET II) in Method 2; and the remaining ten collocations (SET III) in Method 3. The other two groups get the same 30 collocations in the same 3 methods in such a way that no two groups get the same set of collocations in the same method. As can be seen in Table 2, the mean scores of each test method are now the same, as expected.

Table 2. *Hypothetical test scores (counterbalanced, option E)*

	Student #	M1	M2	M3
Group A	1	8	7	6
	2	9	8	7
	3	10	9	8
	4	8	7	6
	5	7	6	5
Group B	6	7	6	8
	7	8	7	9
	8	9	8	10
	9	7	6	8
	10	6	5	7
Group C	11	6	8	7
	12	7	9	8
	13	8	10	9
	14	6	8	7
	15	5	7	6
	Mean	7.4	7.4	7.4

Note: SET I (0); SET II (-1); SET III (-2)

However, one further question needs to be considered: Does counterbalancing remove any possible bias that may be introduced by unequal sets of collocations? In other words, does it matter how large the differences are between the sets? Using the same design as in Table 2, it can be demonstrated that different degrees of disparity between the three sets of collocations have different degrees of influence on the validity of the results of the experiment. In both Table 3 and Table 4, the collocations in SET I are, on average, the easiest, whereas the collocations in SET III are the most difficult. However, the difference between any one set and the next more difficult or easier one can be 1 point (Table 3), or it can be 2 points (Table 4). Note also that in both of these hypothetical situations, it is assumed that Method 1 is the easiest, and Method 3 is the most difficult, so that any student who scores, say, 8 points for 10 collocations in M1 will score 7 points for another 10 collocations in M2, provided that the sets of collocations in both methods are, on average, of the same difficulty. For example, Student 13 (Table 4) scores 6 points for M1 (SET III); now, given that M2 is one point more difficult than M1, and that SET I is four points easier than SET III, the student will score 9 points (6-1+4) for M2 (SET I).

Table 3. Hypothetical test scores (smaller difference between sets)

Student #	M1	M2	M3
	(0)	(-1)	(-2)
1	8	6	4
2	9	7	5
3	10	8	6
4	8	6	4
5	7	5	3
6	7	5	6
7	8	6	7
8	9	7	8
9	7	5	6
10	6	4	5
11	6	7	5
12	7	8	6
13	8	9	7
14	6	7	5
15	5	6	4
Mean	7.4	6.4	5.4

Note: SET I (0); SET II (-1); SET III (-2)

Table 4. Hypothetical test scores (larger difference between sets)

Student #	M1	M2	M3
	(0)	(-1)	(-2)
1	8	5	2
2	9	6	3
3	10	7	4
4	8	5	2
5	7	4	1
6	6	3	6
7	7	4	7
8	8	5	8
9	6	3	6
10	5	2	5
11	4	7	4
12	5	8	5
13	6	9	6
14	4	7	4
15	3	6	3
Mean	*6.4*	*5.4*	*4.4*

Note: SET I (0); SET II (-2); SET III (-4)

Interestingly, thanks to counterbalancing, the mean score difference between M1–M2, and between M2–M3 is, as expected, in both situations the same (1 point), although due to the greater difficulty of the collocations in Table 4, the means are, on average, lower than those in Table 3. Now, using the dependent (or paired-samples) *t*-test, we can compare the means of the test methods in both situations and see whether they are significantly different at $\alpha = .05$. SPSS Output 1 presents the results of this analysis using the data in Table 3, and SPSS Output 2 comes from the same statistical procedure, but applied to the data in Table 4.[21] As can be seen, when the difference between the sets of collocations is relatively small (Table 3), counterbalancing successfully minimizes the bias, and the difference between the test methods is statistically significant (SPSS Output 1). However, when the difference between the sets of collocations is larger (Table 4), the experiment results in a Type II error because it fails to detect a genuine effect (i.e. the assumed 1 point difference between the test methods is no longer significant: see SPSS Output 2).[22]

SPSS Output 1. *Dependent t-test results (smaller difference)*

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	M1 - M2	1.00000	1.46385	.37796	.18935	1.81065	2.646	14	.019
Pair 2	M2 - M3	1.00000	1.46385	.37796	.18935	1.81065	2.646	14	.019

SPSS Output 2. *Dependent t-test results (larger difference)*

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	M1 - M2	1.00000	2.92770	.75593	-.62131	2.62131	1.323	14	.207
Pair 2	M2 - M3	1.00000	2.92770	.75593	-.62131	2.62131	1.323	14	.207

In conclusion, neither randomizing not counterbalancing alone can guarantee an adequate precision of the experimental design, but rather both should be used. Thanks to a random selection of collocations followed by their random assignment to as many sets as the test methods to be compared, and thanks to a counterbalanced administration of the tests, the potentially confounding effects of the collocations' unequal difficulties can be kept to a minimum. In other words, any differences between individual collocations will be spread equally among the sets and among the methods.

---

21 Strictly speaking, a repeated-measures ANOVA would be appropriate prior to conducting the *t*-tests. However, for the sake of simplicity, that first step has been omitted here.

22 The results of the *t*-tests were obtained with SPSS (SPSS for Windows, Release 14.0.0, SPSS Incorporated 2005).

It can be further demonstrated that unequal groups of students may also invalidate the experiment by causing floor or ceiling effects to occur. For example, Table 5 presents the same scores as Table 4, but adjusted to take account of the following differences between the groups: -3 for Group B, and -6 for Group C. The scores are also corrected (in parentheses) by changing every negative number to zero. If language ability could take on negative values, no floor effects would occur, and the differences between the means of the test methods would be the same as in Table 4 (1 point). Moreover, the statistical significance of those differences would be exactly the same as in SPSS Output 2. However, with the negative numbers changed to zeroes, the differences between the means have changed, as has their statistical significance (SPSS Output 3 shows the results of the dependent *t*-test conducted with the corrected data from Table 5). Therefore, in order to avoid floor or ceiling effects, the groups should be balanced in terms of ability, either through randomization or on the basis of the students' previous performance.

Table 5. *Hypothetical test scores (unequal groups)*

	Student #	M1	M2	M3
		(0)	(-1)	(-2)
Group A (0)	1	8	5	2
	2	9	6	3
	3	10	7	4
	4	8	5	2
	5	7	4	1
Group B (-3)	6	3	0	3
	7	4	1	4
	8	5	2	5
	9	3	0	3
	10	2	-1 (0)	2
Group C (-6)	11	-2 (0)	1	-2 (0)
	12	-1 (0)	2	-1 (0)
	13	0	3	0
	14	-2 (0)	1	-2 (0)
	15	-3 (0)	0	-3 (0)
	Mean	3.4 (3.9)	2.4 (2.5)	1.4 (1.9)

Note: *SET I (0);* *SET II (-2);* *SET III (-4)*

SPSS Output 3. *Dependent t-test results (unequal groups)*

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	M1 - M2	1.46667	2.19957	.56793	.24859	2.68475	2.582	14	.022
Pair 2	M2 - M3	.53333	2.61498	.67518	-.91479	1.98146	.790	14	.443

4. Uncontrolled variables and interactions

A possible limitation of the design described above is that it does not explicitly control for the potentially confounding effects of various extraneous variables. By way of illustration, consider a simple counterbalanced design (Table 6) in which two groups of students take two equivalent test forms: Form A consisting of SET I (Method 1) and SET II (Method 2); and Form B consisting of SET II (Method 1) and SET I (Method 2). On the basis of the test scores, we can compare the means of the two methods to see whether they are significantly different.

Table 6. *A simple counterbalanced design*

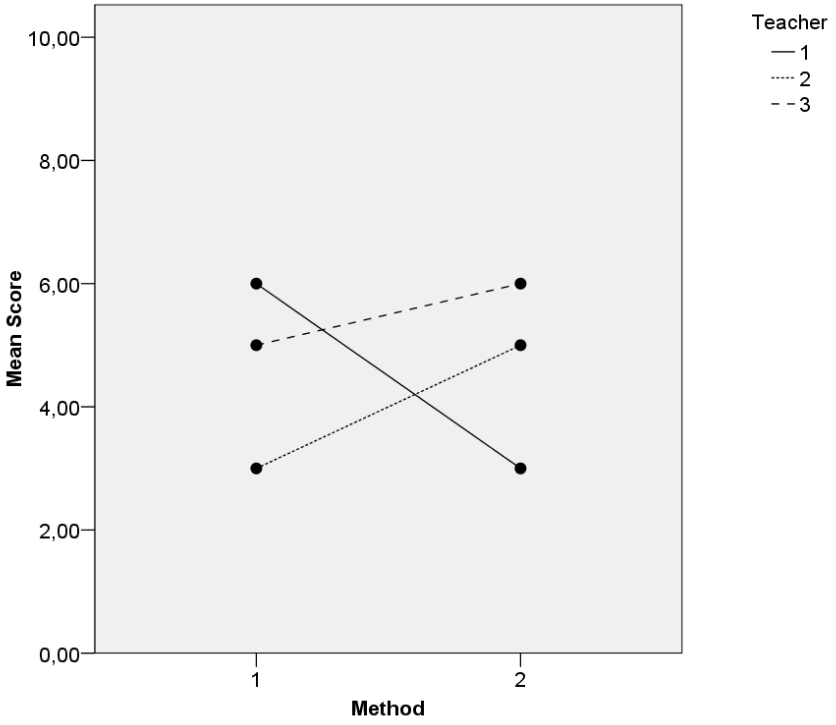
	Students	Method 1	Method 2	
Group A	1 2 3	SET I	SET II	Form A
Group B	4 5 6	SET II	SET I	Form B
		Mean score	*Mean score*	

Fundamental to this experimental design is the assumption that the difficulty of the sets of collocations constitutes a variable that is controlled (by means of random assignment). In other words, its influence on test performance is assumed to be predictable, which means that we expect weak students to do poorly and good students to do well on both sets of collocations. Nevertheless, suppose that for certain students the collocations in SET I are actually more difficult than those in SET II while the reverse holds for some other students. In such a case, the observed difference between the test methods will not be attributable solely to the experimental modification (i.e.

changing the level of the treatment variable). Rather, the outcome of the experiment will be due in no small measure to the effects of miscellaneous extraneous variables such as the students' learning styles, courses attended, books and resources studied, formal instruction, etc.

Furthermore, all of the above-mentioned uncontrolled variables may also interact in some way with the treatment variable. For example, if one teacher happens to place greater emphasis on Method 1 than on Method 2, and two other teachers happen to do the opposite, then the teachers (i.e. instruction) will likely constitute a variable, not controlled in the experiment, which will interact with the test methods, and the combined effect of these variables will impact in an unpredictable way on the students' performance. An example of such an interaction is given in Figure 1: the students taught by Teacher 1 scored, on average, 6 points on Method 1 and 3 points on Method 2, whereas the students taught by Teachers 2 and 3 scored higher on Method 2 than on Method 1.

Figure 1. *An example of a teacher × method interaction*



Therefore, we need to have good reasons to believe that the influence of extraneous variables and their interactions with the treatment variable are not likely to be serious:

- First, while it might be reasonable to expect extraneous variables to interact with

the difficulty of individual collocations, it is less likely that one set of random collocations as a whole will be more difficult for some students than for others, with the reverse being true for another set of collocations. In the research design discussed here, scores on every set of collocations as a whole are used for analysis.

- Second, in a proficiency test administered to a heterogeneous sample of students, the participants' different educational backgrounds would probably impact significantly on their performance in respect of both the collocations and the test methods. On the contrary, in a classroom achievement test which is based on the syllabus, all the students will have had some contact with all of the collocations and with all of the methods.

In sum, the way in which the test scores will be analyzed as well as the relative uniformity of the sample of students are believed to provide adequate safeguards against bias in the experiment.

5. Variables and levels of measurement

Language ability constitutes a continuous variable, even though we report it using discrete values.[23] Just as it is convenient to talk of a length of time in terms of months and days rather than in terms of decimal fractions of a year, so it is convenient to report language proficiency or achievement using distinct units of measurement such as grades, stanines, percentiles, etc. according to the required level of precision. Therefore, although it is theoretically possible to record the achievement of two test takers as .875 and .956 (on a scale from 0 to 1), such a degree of accuracy is not normally necessary. Rather, the achievements might be recorded as, for example, *very good* and *excellent* respectively.

Similarly, although the total score on a test composed of ten dichotomously[24] scored items can take on only eleven distinct values (from 0 to 10), this does not mean that test scores constitute a discrete variable. Rather, the artificially limited variability of what is essentially a continuous variable follows from the measurement procedure. More precisely, it is the consequence of using the narrowest possible (i.e. dichotomous) rating scale in the marking of individual test items.

It is also important to note that the units of measurement used in assessing language ability generally constitute either ordinal or interval scales. Percentile ranks, for example, are only an ordinal scale because the difference in levels of ability between a person who is in the 15th percentile and the one who is in the 25th percentile

23 L. F. Bachman, p. 25.

24 That is, either right (1) or wrong (0).

may not be the same as the difference of 10 percentile ranks at the other end of the score distribution.[25] In the case of test scores, on the other hand, a change from 10 to 20 points corresponds to the same change in levels of ability as a change from 70 to 80 points. Test scores, therefore, constitute interval data.[26]

The level at which variables have been measured is a prime consideration when choosing a suitable statistical procedure to analyze data.[27] For example, parametric tests of significance can only be performed on interval or ratio data. Ordinal data, on the other hand, cannot even be correctly averaged to obtain the arithmetic mean.[28]

6. Concluding remarks

Test method is an important source of test score variance and its impact on test performance should be investigated using carefully designed experiments. While correlational procedures (and most notably the multitrait-multimethod matrix[29]) have traditionally been used in investigations of the method effect in the context of norm-referenced testing, the significance of differences between means seems more important for the purposes of criterion-referenced testing. This should be assessed using procedures such as repeated-measures analysis of variance and/or dependent *t*-tests. The precise choice of statistical test will largely depend on the level of measurement, and, naturally, on the distribution of the data.

25 See, for example, p. 306 *ff.*

26 For a detailed discussion of levels of measurement, see Francuz and Mackiewicz, pp. 29-40.

27 Cf. J. Brzeziński, *Metodologia badań psychologicznych,* Wydanie III, Warszawa 2002, p. 265.

28 See: L. F. Bachman, p. 309.

29 First proposed by D. T. Campbell and D. W. Fiske, *Convergent and discriminant validation by the multitrait-multimethod matrix,* "Psychological Bulletin", 1959 No. 2.