ROCZNIKI HUMANISTYCZNE
Tom LVI, zeszyt 5 – 2008

WOJCIECH MALEC

THE EFFECT OF TASK CHARACTERISTICS ON MASTERY/NON-MASTERY DECISIONS

1. INTRODUCTION

Whatever their particular purposes, language tests are, broadly speaking, constructed and administered with a view to assessing (an aspect of) language ability. Inferences about the ability being measured as well as classification decisions are made on the basis of test takers' scores. However, test score variance is never solely and directly due to variations in language ability. A variety of confounding factors, both external (e.g. weather) and internal (e.g. motivation) can affect test performance. An important source of variance that is not associated with language ability is the method of testing (cf. BACHMAN 2004: 156).

Test method is a general term used to refer to the testing procedure as a whole and as such can be viewed and examined in its entirety. However, within the framework of test method facets, or task characteristics (BACHMAN 1990; BACHMAN and PALMER 1996), 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 can be, for example, the format of the test items and the way it impacts on the difficulty of the test.

It is important to note that test method effects can be of two main kinds. They can be manifested either in [1] different rank orders of the test takers (which means that the methods do not measure the same construct), or in [2] different mastery/non-mastery classifications (which means that the methods impact on test difficulty). It might be claimed that the two kinds of effect are

Dr. WOJCIECH MALEC – Chair of ELT Methodology – English for Students with a Visual Impairment, Department of English Studies, the John Paul II Catholic University of Lublin; address for correspondence: Al. Racławickie 14, PL 20 950 Lublin; e-mail: malew@kul.pl

not equally relevant to both norm-referenced testing (NRT) and criterion-referenced testing (CRT). For example, if Student A scores 70% and Student B scores 80% when one method is used, and they score 40% and 50% respectively when another method is used, the effect of test method is of little interest to NRT. In this approach to testing, it is only an individual's relative standing with respect to the other test takers that matters: as long as there is a significant correlation between the scores obtained from the different test methods, the difference between them can be regarded as negligible. From the perspective of CRT, by contrast, while correlations are also important for construct validation, the difficulty associated with a given test method is of paramount importance. Unless we recognize this kind of effect and learn to understand it, we will not be able to correctly interpret CR test scores. In other words, we will not be able to ascertain whether students do not meet the criterion for mastery because they have not mastered the content domain of the test or because the cognitive demand of the test tasks is very high.

Gaining a better understanding of how item format can impact on the difficulty of a criterion-referenced progress test was the underlying rationale for the empirical study reported in the sections to follow. The measurement instruments were intended to assesses knowledge of collocations, a test construct that is conspicuously under-researched in the measurement literature (cf. BONK 2001). Because of that, one of the purposes of the study was the development of effective measures of collocational knowledge, mainly through quantitative item analysis and through an examination of the tests' validity and reliability (see also MALEC 2006, 2007).

2. PREVIOUS RESEARCH

A number of studies have already demonstrated that test performance can be significantly affected by task characteristics in assessments of speaking (e.g. BACHMAN and PALMER 1982; FULCHER 1996), reading comprehension (e.g. SHOHAMY 1984; ALDERSON and URQUHART 1985; FREEDLE and KOSTIN 1993; M. KOBAYASHI 2002, 2004; W. KOBAYASHI 2005), listening comprehension (e.g. SHOHAMY and INBAR 1991; YING-HUI 2006), and writing (e.g. HAMP-LYONS and PROCHNOW 1991; SPAAN 1993). Within the field of vocabulary testing, a few studies have appeared which throw some light on how test performance can be affected by item format. The findings of HURLBURT (1954), CORRIGAN and UPSHUR (1982) and ARNAUD (1989) all sug-

gest, on the basis of correlations, that depending on the item format (e.g. multiple choice, completion, translation, error correction) as well as on the channel in which the instructions are presented (e.g. aural or visual, printed or picture-cued), the test taps into a different aspect of lexical knowledge.

2.1. *Test format and test difficulty*

Given that there are no studies devoted to the question of item format specifically in the context of collocation testing, and because we may expect, at least to a certain extent, similar method effects across assessments of different constructs, the following discussion draws on research in educational measurement in general. Even a cursory perusal of the literature reveals inconsistencies and conflicting findings concerning the differences in difficulty between selected- and constructed-response formats.

In Ito's (2004) study, test method had a significant effect on test performance, $F(3) = 82.22$, $p < .01$. However, contrary to expectations, MC translation ($M = 19.29$) was not easier than open-ended translation ($M = 20.16$), $t(69) = 1.11$, $p > .05$. Each of these two methods was significantly easier than a cloze test ($M = 11.79$), and significantly more difficult than a short-answer test ($M = 22.50$).

TRUJILLO (2005) found no significant effect of item format on the performance of secondary-English speakers (i.e. those for whom English was a second language), $F(1, 125) = .66$, $p > .05$, on the SAT™ reasoning test. However, for primary-English speakers (i.e. native speakers of English), MC items were easier than constructed-response (gap-filling) items; this effect was statistically significant and substantial, $F(1, 129) = 64.01$, $p < .01$, $\eta^2 = .50$.

O'LEARY (2001) analyzed data from an international mathematics and science test in order to find out whether the rank ordering of countries based on students' test performance was affected by the format of the test. The results showed that MC items were easier than short-answer items, which in turn were easier than extended-response items (cf. *ibid.*, p. 11).

There is also evidence that item format can have a considerable impact on the classification of students. KENNEDY and WALSTAD (1997) compared multiple-choice scores with constructed-response scores obtained from Advanced Placement tests in microeconomics and macroeconomics. If we summarize the differences in classifications, we get the following (for details, see *ibid.*, Table 1, p. 364):

A. In microeconomics, 41% of the students had the same classifications on the basis of MC scores as on the basis of constructed-response scores; 28% would have been classified into a higher category on the basis of MC alone; 31% would have been classified higher if judged solely on their constructed-response scores.
B. In macroeconomics, 42% of the students had the same classifications on the basis of MC scores as on the basis of constructed-response scores; 26% would have been classified into a higher category on the basis of MC alone; 32% would have been classified higher if judged solely on their constructed-response scores.

The above suggests that the effect of item format on test performance may not be the same for all individuals. In language testing, the possibility that such a situation may arise has been pointed out by BACHMAN (1990), who noted that "while one person might perform very well on a multiple-choice test of reading, another may find such tasks very difficult" (p. 113). The findings of the present research were expected to shed some light on the significance of such method × student interaction effects in the context of criterion-referenced measurement, i.e. in situations where mastery/non-mastery classifications are made. The empirical study was in two parts: in the pilot study six item formats were compared, three of which were further analyzed in the main study.

3. METHOD

3.1. *Participants*

The subjects for the experiments were first-year students of English at the John Paul II Catholic University of Lublin. Their proficiency in English ranged from upper-intermediate to advanced. Sixty students (44 females, 16 males), randomly assigned to 6 groups of 10, participated in the pilot study, and fifty one students (37 females, 14 males), divided into 3 groups of 17, took part in the main study. Additionally, forty five upper-intermediate students (38 females, 7 males) from the Teacher Training College in Tomaszów Lubelski participated in pilot testing of the collocation test that was administered in the main study.

3.2. *Materials*

This section presents the most important specifications of the collocation tests: purpose, construct, and test method.

Within READ and CHAPELLE's (2001) framework for second language vocabulary assessment, test *purpose* consists of three components: *inferences* to be drawn from test performance, *uses* of the test results, and *impacts* that the test is intended to have.

In this study, on the basis of test scores, *inferences* were to be made about the test takers' level of mastery of a specific area of language ability (i.e. collocational knowledge).

The collocation tests had two main *uses*. In addition to research uses (the scores were analyzed in order to test experimental hypotheses), the instructional uses involved decision making about the learners, about their achievement of the objectives of the syllabus and about the progress that they had made.

The intended *impact* of the tests was "to encourage the students to study and revise the vocabulary items [and collocations] presented in each unit of their course textbook" (*ibid*., p. 14). These purposes can best be served by discrete, selective, and context-independent measures (see also READ 2000).

The *construct*, or the 'what' of the tests, was defined on the basis of the syllabus that the students were learning. BACHMAN and PALMER (1996) pointed out that "[s]yllabus-based construct definitions distinguish among the specific components of language ability that are included in an instructional syllabus" (p. 118). The specific component which was the focus of assessment was lexical knowledge in general and collocational knowledge in particular. It is impossible to completely separate one from the other. By testing collocations, we are simultaneously testing individual words' form and meaning. In this sense, collocation tests can be said to be measuring the general construct of vocabulary knowledge.

As noted above, for the purposes of measuring students' mastery of the content of a specific lexical syllabus, discrete and selective tests are most appropriate. In addition, they can be relatively context-independent in the sense that test takers do not need to make any inferences about the meaning of each target word (cf. READ and CHAPELLE 2001: 5). SCHMITT (1999) pointed out that "[i]nferencing from context is a valuable skill, but is a different construct from previous vocabulary knowledge" (p. 195). On the other hand, testing collocations in total isolation rather defeats the purpose of measurement. The ideal compromise seems to be testing them in sentence contexts. Support for this comes from HOEY (1991): "each lexical item is stored more or less as received – in the context of the sentence in which it was used" (p. 154). This approach is also very practical: a context sentence that is typical for a given collocation can easily be found or constructed, in

contrast to a longer passage of text. Longer texts can certainly be used in proficiency tests, which do not target specific, pre-selected lexical items taken from, for example, a unit in the coursebook.

In light of the above considerations, test items which purport to measure knowledge of a given collocation should be based on an appropriate context sentence. Accordingly, only those *test methods* (item formats) which are based on a context sentence were chosen for the experiment. More precisely, these were: [1] Fill the Gaps (FG), [2] Multiple Choice (MC), [3] Transformations: Use the Word Given (TW), [4] Transformations: Complete the Sentence (TC), [5] Error Correction (ER), [6] Translation (TR). Such a choice was dictated by practical concerns, too: the students participating in the experiments were fully familiar with these task types. With the exception of TR, all of them had been extensively practised in class prior to the tests. The translation task was included in order to test the quality of each target sentence. It was actually a copy of the gap-filling task in which a Polish translation of the target collocation was additionally provided. The scores obtained from these two item formats were expected to give an indication of whether the context sentences were enough in themselves to elicit the missing target words.

The tests were constructed in such a way that each of the 30 collocations selected for testing was elicited in every item format. This is illustrated in Table 1.

Table 1. Test items for the collocation "couldn't be bothered"

Fill the Gaps (FG)
It was getting dark, but he _____ *couldn't be* _____ bothered to put the lights on.
Multiple Choice (MC)
It was getting dark, but he ____ bothered to put the lights on.
A wouldn't have B didn't feel C wasn't D couldn't be
Transformations: Use the Word Given (TW)
It was getting dark, but he was too lazy to put the lights on. BOTHERED
It was getting dark, but he couldn't be bothered to put the lights on.
Transformations: Complete the Sentence (TC)
It was getting dark, but he was too lazy to put the lights on.
It was getting dark, but he couldn't be bothered to put the lights on.
Error Correction (ER)
It was getting dark, but he ~~wasn't~~ COULDN'T BE bothered to put the lights on.
Translation (TR)
It was getting dark, but *he* _____ *couldn't be* _____ *bothered* to put the lights on.
(*nie chciało mu się*)

The following general guidelines were followed while constructing the items: First, MC items were constructed in accordance with the relevant, widely-accepted principles of item writing (see, for example, HALADYNA *et al.* 2002). The FG item was produced simply by replacing the target word(s) in the collocation (the same as in the MC item) with a blank. For TW and TC, the collocation was paraphrased and the resulting new sentence was part of the prompt used for eliciting the target sentence. In the case of TW, one word from the collocation other than the target word(s) was the only element of the target sentence that was explicitly given; in this sense, the TW format required the longest response. In the case of TC, a part of the target sentence, either to the left or to the right of the target word(s), was given, and the remainder of the target sentence was replaced with a blank. For the ER item, one of the corresponding distractors from the MC item was substituted for the target word(s) in the collocation. The TR item was written by copying the FG item and by adding a Polish translation of a part of the sentence containing the collocation; the whole of the corresponding English expression was italicised.

In this way, six equivalent test forms were constructed, each comprising all of the collocations and all of the item formats (see Table 2 in the next section). Moreover, irrespective of the item format, every given collocation was elicited in the same context sentence, and the general test task involved reconstructing the target sentence. In this sense, it was common to every item format.

3.3. *Procedures and scoring*

A repeated-measures design was used for this study. Besides requiring fewer participants, such experimental designs minimize the probability of making a Type II error (i.e. failing to detect an effect that does genuinely exist). This is because individual differences between participants are controlled, which results in a reduction of unsystematic variability in scores and, by the same token, in greater statistical power to detect an effect (FIELD 2005). On the other hand, a serious drawback to repeated-measures designs is that they create the possibility of transfer between treatment conditions. Therefore, in order to eliminate any potential carryover effects, the design was modified with the help of a Latin square (see Table 2).

Table 2. Experimental design (pilot study)

	F_1	F_2	F_3	F_4	F_5	F_6	
G_1	C_1	C_2	C_3	C_4	C_5	C_6	(A)
G_2	C_6	C_1	C_2	C_3	C_4	C_5	(B)
G_3	C_5	C_6	C_1	C_2	C_3	C_4	(C)
G_4	C_4	C_5	C_6	C_1	C_2	C_3	(D)
G_5	C_3	C_4	C_5	C_6	C_1	C_2	(E)
G_6	C_2	C_3	C_4	C_5	C_6	C_1	(F)

$F_{1\text{-}6}$ – item format
$G_{1\text{-}6}$ – group of students
$C_{1\text{-}6}$ – set of collocations
A-F – test form

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 test methods, the potentially confounding effects of the differences in difficulty between individual collocations could be kept to a minimum. In other words, any differences between individual collocations were to be spread equally among the sets and among the methods.

In order to reduce the artificiality of the experimental situation, the students were given the test forms as part of their regular Practical English tests (called "big tests"). These were administered in a big room and at the same time for all of the students, which minimized the possibility of violating the independence of observations assumption.

The tests were all marked by one person, the researcher, and to further ensure intra-rater reliability, they were re-marked. In the course of the main study, three scoring procedures were compared: [1] SP–*one* (answers containing minor errors, e.g. grammatical ones, received full credit); [2] SP–*half* (answers containing minor errors received partial credit); [3] SP–*zero* (answers containing minor errors received zero credit). The tests were scored by partial credit using SP–*half*. In order to compare the three scoring procedures, two additional sets of data were obtained simply by transforming the scores based on SP–*half* in such a way that every half mark was changed to one mark (SP–*one*) and to zero (SP–*zero*). Naturally, the transformations could not be applied to scores on multiple-choice items, which were only dichotomously scored as either right (1) or wrong (0).

4. RESULTS[1]

4.1. *Item analysis*

In addition to item facility (IF), a traditional NRT statistic, two cut-score indices were calculated for the collocation test items: item phi (ϕ) and the agreement statistic (*A*). Both indices indicate how a given item discriminates at the pass mark, i.e. at "the observed score that corresponds to the domain score associated with mastery level" (BACHMAN 2004: 198).

In the case of the pilot study, the values of the agreement statistic indicated that those students who answered a given item correctly were to a high degree the same as those who passed the test, and those students who answered a given item incorrectly were mostly the same as those who failed the whole test. On the other hand, the value of the item phi was in a few cases below zero, which indicated that the correlation between item and test performance was sometimes negative. Such a situation occurs when, for example, an item is answered incorrectly by those few students who passed the test. By way of illustration, Coll. 2 in SET V (Form E) was answered correctly by 3 students, and incorrectly by 7 students (IF = .30). Of all the ten students in Group 5, only one scored above the pass mark for the whole test. The agreement statistic was relatively high (*A* = .60) because six of the students who answered the item incorrectly also failed the test. However, the item phi was below zero (ϕ = -22) because the only student in this group who passed the test actually answered the item incorrectly. Changing the pass mark to, say, 50% would have resulted in more acceptable values of the cut-score indices, but doing this *ex post facto* was unjustifiable.

In fact, it is not unusual for the values of the agreement statistic to be quite different from the values of the item phi, as noted by BROWN and HUDSON (2002: 124). Furthermore, they suggest that "items should not be rejected solely on the basis of these values" (p. 127), but rather their content should be carefully examined.

When considering whether a given item from the test used in the pilot study was a good discriminator or not, it is worth remembering that the six groups of test takers were relatively small, and achieving high consistency in the scores was not easy. For example, in Form E, the second collocation of SET II (*the crux on the matter*) had an extremely low value of the item phi (ϕ = -.51) mainly because the student who scored highest on the test actually

[1] See MALEC (2006) for details.

answered **the crux of the count*. If only this particular student had answered the item correctly, the value of the index in question would have risen to an acceptable positive value of .17.

As for the main study, the correlation between students' performance on individual items and their performance on the test as a whole was quite acceptable. Most of the items discriminated very well between masters and non-masters, as indicated by the values of *A* above .50, and by positive values of ϕ.

4.2. *Test qualities*

The dependability[2] of test scores as indicators of domain scores is analogous to the concept of internal consistency in NRT. The relevant estimate, called the phi coefficient (Φ), was computed using a short-cut formula derived by BROWN (1990). In order to estimate the consistency of mastery/non-mastery decisions, two squared-error loss agreement indices were calculated for each form of the collocation tests: phi lambda (Φ_λ) and kappa squared (κ^2) (BROWN and HUDSON 2002).

In the pilot study, despite a considerable amount of internal inconsistency in the scores indicated by relatively low values of the Φ coefficient (from .58 to .85), all six forms of the collocation test were highly dependable in terms of mastery/non-mastery decisions (both Φ_λ and κ^2 were greater than .90). In the main study, all of the three coefficients had high values.

Evidence in support of construct validity for the collocation tests was sought in correlations with the lexis and grammar components of the big tests and from principal component analysis.

In the pilot study, the Pearson product-moment correlation coefficients were calculated for each of the following pairs: collocations–lexis, collocations–grammar, and lexis–grammar. Before that, the data were checked for normality of distribution. For every set of data, neither the Kolmogorov-Smirnov test nor the Shapiro-Wilk test was significant. As can be seen in Table 3, the correlations between collocations and lexis were the highest for every group of students, which was an indication of convergent and divergent validity. The item formats on the lexis and grammar tests were similar to those on the collocation test (mostly completion and transformation types), so the influence of method variance was assumed to be insignificant.

[2] In the context of CRT, the term 'dependability' is often used instead of 'reliability'.

In short, the pattern of correlations was interpreted to indicate that the collocation test forms measured a construct that was very similar to the construct measured by the lexis test and somewhat different from the construct measured by the grammar test.

Table 3. Intercorrelations between collocation, lexis, and grammar scores

		Lexis	Grammar
Form A	Collocations	.929**	.808**
	Lexis		.759*
Form B	Collocations	.754*	.383
	Lexis		.575
Form C	Collocations	.951**	.580
	Lexis		.690*
Form D	Collocations	.913**	.741*
	Lexis		.662*
Form E	Collocations	.911**	.496
	Lexis		.457
Form F	Collocations	.854**	.793**
	Lexis		.710*

** Correlation is significant at the .01 level (2-tailed)

* Correlation is significant at the .05 level (2-tailed)

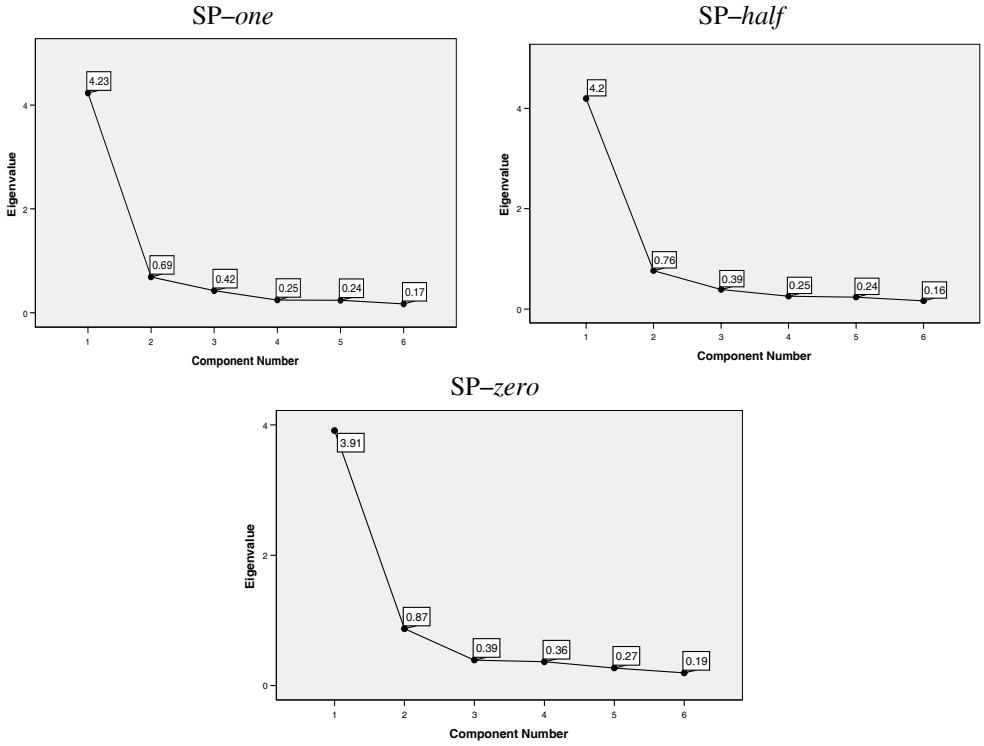
A very similar pattern of correlations was observed in the main study. However, because of the fact that the scores were this time significantly non-normal, as revealed by the Shapiro-Wilk test, a non-parametric correlation coefficient was calculated. Specifically, Kendall's tau (τ) was chosen because the data sets were relatively small and there were quite a few tied ranks in the collocation data (cf. FIELD 2005: 131). The values of the correlation coefficients indicated that collocational knowledge emerged on that test as a construct that was more similar to lexis than to grammar, as expected.

Finally, the unidimensionality of the collocation tests was supported by principal component analysis (PCA). Although several factors were extracted in this analysis, there was only one component with an eigenvalue greater than 1. As noted by BROWN and HUDSON (2002: 205), if one factor

is clearly dominant, this points to unidimensionality. A test that is unidimensional is believed to measure one single trait, or ability.

The results of PCA (main study) presented diagrammatically in Figure 1 indicate that the degree of unidimensionality of the test was different depending on which scoring procedure was used. The difference between the first two eigenvalues was largest in the case of SP–*one* and smallest in the case of SP–*zero*. The ratio of the first eigenvalue to the second was 6.13 for SP–*one*; 5.53 for SP–*half*; and 4.49 for SP–*zero*. The same pattern of differences was observed for the values of the communalities, total variance explained, and factor loadings. In general, the fit of the model was best in the case of SP–*one*, as confirmed by the smallest number of nonredundant residuals (53%) with absolute values greater than .05.

Figure 1. Scree plots of PCA results with eigenvalues



The outcome of this analysis can be interpreted as being indicative of construct validity for the second collocation test. The argument is as follows.

If grammar and spelling were not different constructs from collocational knowledge, then their influence on test score variance would not have affected the unidimensionality of the test. In other words, the test would have appeared to measure only one trait or ability to the same degree with and without the influence of grammar and spelling. The scoring procedures differed from each other precisely in the way grammar and spelling errors were treated. Under SP–*one*, such errors were ignored completely; under SP–*half*, their influence was moderate; and under SP–*zero*, they had the same status as lexical errors. The fact that the scores based on SP–*one* were characterized by the highest degree of unidimensionality is clear evidence that grammar and spelling on the one hand and collocations on the other do not constitute the same construct.

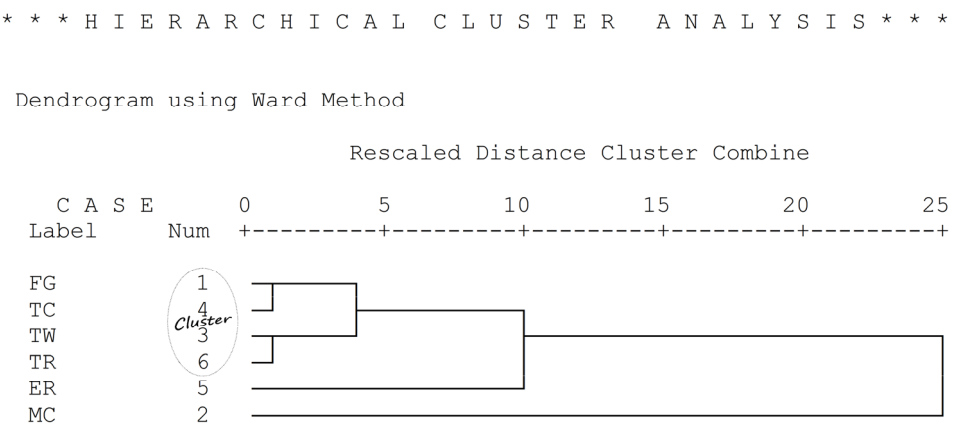
4.3. *Item format comparisons*

A repeated-measures analysis of variance was performed despite the fact that not all of the data were normally distributed. The reason for this was that ANOVA is very robust to violations of the assumption of normality (cf. FRANCUZ and MACKIEWICZ 2005). However, non-parametric tests were also conducted in order to see whether they would produce comparable results (main study only).

In the pilot study, ANOVA showed that item format had a significant effect on test performance, $F(5, 295) = 21.49$, $p < .001$, $\omega = .39$. *Post hoc* tests (with confidence intervals and significance adjusted using Bonferroni's correction) showed that MC and ER were significantly different from all the other item formats. The differences between FG, TW, TC, and TR were not statistically significant.

The results of ANOVA were similar to the solutions obtained from hierarchical cluster analysis. The default measure of the degree of similarity (squared Euclidean distance) was used in this analysis. Ward's method of creating clusters was chosen because it accentuates the differences between cases (FRANCUZ, personal communication, 2006). Prior to the analysis, the data were standardized by converting to Z-scores, as recommended by FIELD (2000: 8).

Figure 2. *Results of hierarchical cluster analysis*



Cluster Membership

Case	5 Clusters	4 Clusters	3 Clusters	2 Clusters
Fill the Gaps	1	1	1	1
Multiple Choice	2	2	2	2
Transformations: Use the Word Given	3	3	1	1
Transformations: Complete the Sentence	1	1	1	1
Error Correction	4	4	3	1
Translation	5	3	1	1

The dendrogram in Figure 2 indicates that FG and TC as well as TW and TR appeared to be the most similar item formats. These four were then joined to form one cluster. On the other hand, ER and MC emerged as visibly different from the others. The cluster membership analysis clearly shows that in the two-cluster solution, MC stood out from all of the other item formats. In the three-cluster solution, ER arose as a different entity, in addition to MC. In the four- and five-cluster solutions, the differences were relatively small.

As for the main study, in which MC, TW, and ER were compared, there was a significant main effect of item format on test scores: $F(1.61, 80.65) = 77.63$, $p < .001$ (SP–*one*); $F(1.57, 78.32) = 96.55$, $p < .001$ (SP–*half*); $F(1.68, 84.21) = 104.40$, $p < .001$ (SP–*zero*). Planned comparisons revealed that every item format was significantly different from every other one. Moreover, each comparison yielded a very large effect size. Friedman's ANOVA and Wilcoxon signed-rank tests with a Bonferroni correction for the number of tests produced similar results.

It was not the case that all students without exception scored higher on TW than on ER and still higher on MC. Departures from the predominant pattern can easily be found in Table 4 (to save space, only a portion of the original table is given here). Nevertheless, as far as the mastery/non-mastery decisions are concerned, there is only one example out of 153 cases (51 students multiplied by 3 scoring procedures) of a student being classified as a master on the basis of a more difficult item format and a non-master on the basis of an easier one (No. 12, SP–*zero*, shaded). The overall differences between the three item formats and the scoring procedures expressed in terms of the number of students classified as masters are summarized at the bottom of the table.

Table 4. Mastery/non-mastery decisions

Maximum score: 10; Cut score: 6; Decision: master (+) or non-master (-)

ID	SP–*one*						SP–*half*						SP–*zero*					
	Score			Decision			Score			Decision			Score			Decision		
	MC	TW	ER	MC	TW	ER	MC	TW	ER	MC	TW	ER	MC	TW	ER	MC	TW	ER
1	10	10	9	+	+	+	10	10	8.5	+	+	+	10	10	8	+	+	+
2	10	9	5	+	+	-	10	8.5	4.5	+	+	-	10	8	4	+	+	-
3	10	8	7	+	+	+	10	7.5	7	+	+	+	10	7	7	+	+	+
4	9	10	8	+	+	+	9	9.5	7	+	+	+	9	9	6	+	+	+
5	9	10	9	+	+	+	9	10	9	+	+	+	9	10	9	+	+	+
6	10	9	9	+	+	+	10	8	8.5	+	+	+	10	7	8	+	+	+
7	10	10	8	+	+	+	10	9.5	7	+	+	+	10	9	6	+	+	+
8	10	10	5	+	+	-	10	9.5	5	+	+	-	10	9	5	+	+	-
9	10	9	6	+	+	+	10	9	5.5	+	+	-	10	9	5	+	+	-
10	10	9	9	+	+	+	10	9	8.5	+	+	+	10	9	8	+	+	+
11	10	5	4	+	-	-	10	4.5	4	+	-	-	10	4	4	+	-	-
12	9	8	6	+	+	+	9	6.5	6	+	+	+	9	5	6	+	-	+
13	9	7	3	+	+	-	9	6	3	+	+	-	9	5	3	+	-	-
14	10	8	8	+	+	+	10	7.5	7.5	+	+	+	10	7	7	+	+	+
15	7	2	1	+	-	-	7	1.5	0.5	+	-	-	7	1	0	+	-	-
16	10	8	5	+	+	-	10	7	3.5	+	+	-	10	6	2	+	+	-
17	6	3	2	+	-	-	6	3	2	+	-	-	6	3	2	+	-	-
(...)	(...)						(...)						(...)					
50	6	3	2	+	-	-	6	2.5	2	+	-	-	6	2	2	+	-	-
51	1	0	0	-	-	-	1	0	0	-	-	-	1	0	0	-	-	-
Number of masters:				50	42	27				50	39	26				50	33	26

The differences between the three item formats are very large indeed: for example, only one student would have failed on the basis of MC alone, whereas almost every second student would have failed on the basis of ER alone. As for the influence of the scoring procedure, the difference between

MC and TW is smaller under SP–*one* than under SP–*zero* while the opposite is true for the difference between TW and ER.

5. General discussion and conclusions

This study demonstrates that the effect of task characteristics can be investigated using a very economical experimental design, one which, contrary to traditional approaches, requires only a single administration of the measurement instrument. This counterbalanced, Latin-square-based design has important methodological advantages in that it eliminates the confounding effects of many variables. Most importantly, effects of order and practice are removed completely.

A potential weakness of this design is its reliance on the absence of collocation × student as well as method × student interaction effects. However, the results of this study give reasons to believe that the effects of these interactions may be negligible (cf. Table 4). First, while it might be reasonable to expect extraneous variables (e.g. instruction, personal preferences) to interact with the difficulty of individual collocations, it is less likely that one set of random collocations as a whole should be more difficult for some students than for others, with the reverse being true for another set of collocations. In this experimental design, scores on every set of collocations as a whole were 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. By contrast, 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 test scores were analyzed as well as the relative uniformity of the sample of students seem to have provided adequate safeguards against bias in the experiment. However, it must be added that more research is needed into the nature of such interactions in the context of classroom testing.

In the course of statistical analyses, multiple choice came out as the easiest item format, and error correction as the most difficult one. Fill the gaps, translation, and both types of transformations were similar in terms of difficulty. All of the item formats were fairly unidimensional: they measured a similar construct, one that is close to general lexical knowledge and somewhat different from grammatical knowledge. Therefore, it seems reasonable to conclude

that they do not rank-order test takers in significantly different ways. On the other hand, their varying difficulties have an impact on whether a student is classified as a master or a non-master on the basis of test scores. Because of the fact that this kind of effect has a direct bearing on classification decisions, it is particularly significant in the context of criterion-referenced testing.

The most logical interpretation of this effect seems to be that FG, TW, TC, and TR items tap into a different aspect of lexical or collocational knowledge than do MC items on the one hand and ER items on the other. More precisely, the easiness associated with the MC format most probably stems from its lowest cognitive demand, i.e. from the fact that successful performance on items in this format requires only receptive knowledge of the target collocations, in contrast to items in the other formats, all of which require test takers to actually produce (a portion of) the target collocations. It is generally accepted that productive (active) knowledge is a subset of receptive (passive) knowledge.[3] Therefore, if test takers can successfully supply the missing target word, it follows that they should also be able to select it from among the choices, no matter how plausible the distractors are.

Nevertheless, as mentioned earlier, statements can be found in the literature which suggest that it is not entirely obvious that for all individuals selected-response items are easier than constructed-response items:

> The characteristics of the testing method and administration procedure will have a *systematic* effect on test scores, since they may affect different individuals differently. Some test takers, for example, prefer multiple-choice tests, and do better on these, while others perform better on tests that require them to respond in writing.
>
> (BACHMAN 1990: 156, italics in original)

If this were true, however, it would be logical to expect a group of test takers to perform similarly, on average, both on selected- and on constructed-response items: lower scores of some of the testees should be offset by higher scores of some other testees irrespective of the response format so that the means of both formats should be, more or less, the same. The results of this study do not indicate that this might be the case. It is probably a popular misconception, particularly among the students themselves, that constructed-response items may be easier. Not uncommonly, students can be heard to say that they indeed prefer to give an extended response, either

[3] E.g. ARNAUD and SAVIGNON (1997: 158), although MONDRIA and WIERSMA (2004) found that "productive knowledge does not in all cases include receptive knowledge, as is often assumed" (p. 96).

written or oral, rather than take a multiple-choice test, but this should not be taken to mean that they would actually perform worse on MC items, assuming of course that the content would be exactly the same. It is not unlikely that students who say they prefer extended written responses simply hope to be able to get off the subject with impunity, and that those who prefer oral exams simply count on the examiner putting them on the right track. Furthermore, when students know that they are much better at writing or speaking than other students, they might assume that these skills will contribute significantly to the result of the exam, over and above the knowledge itself that is the focus of assessment. And finally, they may feel that constructed-response items give them more flexibility, in the sense that they are free to give any answer that makes sense, not necessarily the keyed response. However, the fact that certain constructed-response items are incapable of eliciting the expected responses cannot be used to argue that they are intrinsically easier than selected-response items.

The experiments in this study were designed in such a way that items in every response format covered precisely the same content, which allows us to determine with a fair amount of certainty which tasks are indeed easier than others, at least as far as the assessment of lexical/collocational knowledge is concerned. As mentioned above, it is most probably the distinction between receptive and productive knowledge that is behind the differences in difficulty of the six item formats. To be more precise, an aspect of reception referred to as *recognition* is believed to be measured by MC items, and an aspect of production known as *recall* is apparently measured by FG, TW, TC, and TR items (see READ 2000: 154 *et seq*. for a discussion of a twofold distinction between receptive and productive knowledge). In the case of the collocation tests, recognition items presented test takers with four choices and required them to identify the word(s) that correctly fitted the given sentential context. Recall items, on the other hand, provided test takers "with some stimulus designed to elicit the target word(s) from their memory" (*ibid*., p. 155).

The only item format that does not readily fit this pattern is error correction, although in PARIBAKHT and WESCHE (1997: 184), "[f]inding the mistake […] in a sentence and correcting it" is given as an example of production exercises. While an aspect of production is certainly present in ER items, they must also involve something else. It is tempting to suggest that what they really require is a combination of recognition and recall because test takers need to identify the wrong word(s) (recognition) and then supply the correct one(s) (recall). At first blush, this line of argument makes it diffi-

cult to reconcile the results of the study with the general consensus that production is subsumed under reception, i.e. that productive knowledge presupposes receptive knowledge. In other words, an additional element of recognition should not increase the difficulty of items which require productive knowledge anyway. Upon closer inspection, however, it *is* possible to accept the argument that ER items test both recognition and recall as long as we acknowledge that recognizing *correct* language does not call for the same type of knowledge or ability as recognizing *incorrect* language. For example, while it may be relatively easy to recognize 'put sb at risk' as a correct collocation, identifying 'put sb at danger' as incorrect is a more demanding task. This difference in difficulty is especially evident when we begin to learn a new language: recognizing an expression that we have memorized is quite straightforward, but we have very little idea of whether some other similar expression is correct or not. What is more, when incorrect collocations are presented in the context of a sentence, the difficulty of ER items is compounded by the fact that test takers do not know exactly where to look for the mistakes. It is, therefore, unsurprising that the cognitive demand of items in the error-correction format is the highest.

REFERENCES

ALDERSON, J. C. and A. H. Urquhart. (1985). The effect of students' academic discipline on their performance on ESP reading tests. *Language Testing* 2, 192–204.

ARNAUD, P. (1989). Vocabulary and grammar: a multitrait-multimethod investigation. *AILA Review* 6, 56–65.

ARNAUD, P. and S. J. SAVIGNON. (1997). Rare words, complex lexical units and the advanced learner. In: J. COADY and T. HUCKIN (eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 157-173). Cambridge: Cambridge University Press.

BACHMAN, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

BACHMAN, L. F. (2004). *Statistical Analyses for Language Assessment*. Cambridge: Cambridge University Press.

BACHMAN, L. F. and A. S. PALMER (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly* 16, 449–465.

BACHMAN, L. F. and A. S. PALMER (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

BONK, W. (2001). Testing ESL learners' knowledge of collocations. In: T. HUDSON and J. D. BROWN (eds.), *A Focus on Language Test Development: Expanding the Language Proficiency Construct across a Variety of Tests* (pp. 113-142). Honolulu: University of Hawai'i Second Language Teaching and Curriculum Center.

BROWN, J. D. (1990). Short-cut estimates of criterion-referenced test consistency. *Language Testing* 7, 77-97.

BROWN, J. D. and T. HUDSON (2002). *Criterion-referenced Language Testing*. Cambridge: Cambridge University Press.

CORRIGAN, A. and J. A. UPSHUR (1982). Test method and linguistic factors in foreign language tests. *Iral* 20, 313–321.

FIELD, A. (2000). Postgraduate statistics: cluster analysis. [Online at: http://www.sussex.ac.uk/users/andyf/cluster.pdf]

FIELD, A. (2005). *Discovering Statistics Using SPSS* (Second edition). London: Sage Publications.

FRANCUZ, P. and R. MACKIEWICZ R. (2005). *Liczby nie wiedzą, skąd pochodzą: przewodnik po metodologii i statystyce nie tylko dla psychologów*. Lublin: Wydawnictwo KUL.

FREEDLE, R. and I. KOSTIN (1993). *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: Educational Testing Service.

FULCHER, G. (1996). Testing tasks: issues in task design and the group oral. *Language Testing* 13, 23-51.

HALADYNA, T. M., S. M. DOWNING, and M. C. RODRIGUEZ (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education* 15, 309-334.

HAMP-LYONS, L. and S. PROCHNOW (1991). Prompt difficulty, task type, and performance. In: S. ANIVAN (ed.), *Current Developments in Language Testing* (pp. 58-76). Singapore: SEAMEO Regional Language Centre.

HOEY, M. (1991). *Patterns of Lexis in Text*. Oxford: Oxford University Press.

HURLBURT, D. (1954). The relative value of recall and recognition techniques for measuring precise knowledge of word meanings: nouns, verbs, adjectives. *Journal of Educational Research* 47, 561-576.

ITO, A. (2004). Two types of translation tests: their reliability and validity. *System* 32, 395-405.

KENNEDY, P. and W. B. WALSTAD (1997). Combining multiple-choice and constructed-response test scores: an economist's view. *Applied Measurement in Education* 10, 359-375.

KOBAYASHI, M. (2002). Method effects on reading comprehension test performance: text organization and response format. *Language Testing* 19, 193-220.

KOBAYASHI, M. (2004). Investigation of test method effects: text organization and response format: a response to Chen, 2004. *Language Testing* 21, 235-244.

KOBAYASHI, W. (2005). *An investigation of method effects on reading comprehension test performance*. Paper presented at the 4th Annual JALT Pan-SIG Conference, Tokyo, Japan, 14-15 May 2005.

MALEC, W. (2006). *The Impact of Item Format on Test Performance in Criterion-referenced Assessment of Collocations.* Unpublished PhD thesis, KUL, Lublin.

MALEC, W. (2007). Efekt metody w pomiarze sprawdzającym na przykładzie testowania kolokacji języka angielskiego. In: B. NIEMIERKO and M. K. SZMIGEL (eds.), *Uczenie się i egzamin w oczach uczniów* (pp. 305-315). Kraków: Grupa Tomami.

MONDRIA, J. A. and B. WIERSMA (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: what difference does it make? In: P. BOGAARDS and B. LAUFER (eds.). *Vocabulary in a Second Language: Selection, Acquisition and Testing* (pp. 79-100). Amsterdam: John Benjamins Publishing Company.

O'LEARY, M. (2001). *Item format as a factor affecting the relative standing of countries in the Third International Mathematics and Science Study (TIMSS).* Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

PARIBAKHT, T. S. and M. WESCHE (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In: J. COADY and T. HUCKIN (eds.), *Second Language Vocabulary Acquisition: A Rationale for Pedagogy* (pp. 174-200). Cambridge: Cambridge University Press.

READ, J. (2000). *Assessing Vocabulary*. Cambridge: Cambridge University Press.

READ, J. and C. A. CHAPELLE (2001). A framework for second language vocabulary assessment. *Language Testing* 18, 1-32.

SCHMITT, N. (1999). The relationship between TOEFL vocabulary items and meaning, association, collocation and word-class knowledge. *Language Testing* 16, 189-216.

SHOHAMY, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing* 1, 147-170.

SHOHAMY, E. and O. INBAR (1991). Construct validation of listening comprehension tests: the effect of text and question type. *Language Testing* 8, 23-40.

SPAAN, M. (1993). The effect of the prompt in essay examinations. In: D. DOUGLAS and C. CHAPELLE (eds.). *A New Decade of Language Testing Research* (pp. 98-122). Alexandria, VA: TESOL Publications.

TRUJILLO, J. L. (2005). *The Effect of Format and Language on the Observed Scores of Secondary-English Speakers.* Unpublished PhD thesis, Florida State University, Tallahassee.

YING-HUI, H. (2006). An investigation into the task features affecting ELF listening comprehension test performance. *The Asian EFL Journal Quarterly* 8(2), 33-54.

WPŁYW WŁAŚCIWOŚCI ZADANIA TESTOWEGO NA DECYZJE KLASYFIKACYJNE

Streszczenie

Artykuł jest opisem badania, którego celem było udzielenie odpowiedzi na pytanie, czy właściwości zadania testowego mają wpływ na decyzje klasyfikacyjne w kontekście pomiaru sprawdzającego. Zastosowane narzędzia pomiaru, które sprawdzały znajomość kolokacji (łączliwości wyrazów) języka angielskiego, składały się zarówno z zadań zamkniętych, jak i otwartych różnego typu. Wszystkie testy kolokacji charakteryzowały się wysokim poziomem rzetelności decyzji klasyfikacyjnych oraz trafnością teoretyczną. Analiza wariancji z powtarzanym pomiarem wykazała istotność efektu metody testowania, polegającego na tym, że osiągnięcie przez studenta progu zaliczeniowego zależy od zastosowanej formy zadania testowego.

Streścił Wojciech Malec

Słowa kluczowe: pomiar sprawdzający, efekt metody, aspekty metody testowania, forma zadania, decyzje klasyfikacyjne, testowanie kolokacji.

Key words: criterion-referenced measurement, method effect, task characteristics (test method facets), item format, classification decisions, collocation testing.