ROCZNIKI HUMANISTYCZNE
Tom LXIX, zeszyt 5 – 2021
DOI: http://doi.org/10.18290/rh21695-15

MAŁGORZATA KRZEMIŃSKA-ADAMEK
WOJCIECH MALEC

THE DEVELOPMENT OF AN EFL TEST FOR THE *YOUNGSTER* PROJECT IN POLAND

Abstract. This article is a report on the development of an online assessment instrument, namely a test of English as a foreign language intended for the participants of the Youngster project in Poland. The project has been launched with a view to providing free English education for learners in rural areas. One of the requirements of the project is the provision of effective assessment procedures, regularly conducted at three different points in time — the beginning of the school year, the end of the first semester, and the end of the school year. The purpose of the study was therefore to develop an instrument which could serve both placement and progress functions within the project. This article describes the activities performed prior to the test's operational administration, and focuses primarily on the development of multiple-choice items.

Keywords: EFL test development; multiple-choice format; item analysis; distractor evaluation.

INTRODUCTION

This paper reports on the development of a new assessment tool – a test of English as a foreign language for the participants of the *Youngster* project, which offers free English courses for pre-high school learners in rural areas in Poland. The project was launched in 2007 under the auspices of the European Fund for the Development of Polish Villages with a view to evening out educational opportunities for students who wish to continue their education in high schools of their choice. Participation in additional English

Dr. MAŁGORZATA KRZEMIŃSKA-ADAMEK – Maria Curie-Skłodowska University, Faculty of Humanities, Department of English and General Linguistics; address for correspondence: Plac Marii Curie-Skłodowskiej 4A, 20-031 Lublin; e-mail: malgorzata.krzeminska-adamek@umcs.pl; ORCID: https://orcid.org/0000-0003-2461-2397.

Dr. habil. WOJCIECH MALEC – John Paul II Catholic University of Lublin, Institute of Linguistics, Department of Theoretical Linguistics; address for correspondence: al. Racławickie 14 20-950 Lublin; e-mail: wojciech.malec@kul.pl; ORCID: https://orcid.org/0000-0002-6944-8044.

classes within the project is voluntary. The general idea behind the programme is to provide an extension to the standard English curriculum, therefore the learning content as well as the materials employed differ from those used during regular classes in schools. In order to achieve a high quality of language education, students are organised to work in small groups with the assistance of teachers participating in regular training sessions and utilising effective, up-to-date teaching methods.

Notwithstanding the rather informal character of the courses, the program requires effective assessment procedures to collect evidence that the assumptions of the project are adequately met. The role of assessment within the programme is actually twofold. In the first place, the tests perform a placement function in that they provide information about the current level of the learners' proficiency and thus aid in the formation of study groups. Additionally, the results of placement tests are helpful in the coursebook selection process. Secondly, the tests are to verify whether progress in learning the language is taking place. The feedback obtained from the tests can be used by students and teachers in order to pinpoint specific areas of difficulty and, consequently, to maximise opportunities for learning.

The need for assessing the participants' language abilities has been acknowledged by all project stakeholders, including the European Fund, the teachers, and the partners of the project, who have been responsible for providing methodological support and teaching materials. There were some initial assumptions made about the test to be designed. First of all, it was to be well aligned with the coursebooks used in the project, i.e. it was to constitute a representative sample of the language areas covered in the teaching materials. Furthermore, the tests were to reflect the students' level of preparation for the school-leaving exam. Although exam preparation as such was beyond the scope of the project, the students' success in the exam could potentially contribute to improving their educational opportunities overall, which fits well with the assumptions of the European Fund.

As far as the type of tasks used in the test is concerned, the new instrument was to consist solely of multiple-choice (MC) items. This decision was taken on the grounds that MC items are used very extensively in both small-scale and large-scale foreign language testing, at various levels of proficiency. This item format has been found versatile enough to test diverse areas of knowledge and competence, such as listening (Koyama, Sun, and Ockey), reading (Davison et al.), vocabulary (Gyllstad, Vilkaitė, and Schmitt), collocations (Sonbul and Schmitt), or grammar (Frizelle, O'Neill, and Bishop). Additionally, in terms of practicality of use, MC items have been applied widely because of their ease of administration and the possibility of objective scoring.

The process of pre-operational testing (Kenyon and MacGregor), which is the object of the study reported in this paper, is aimed at assessing the appropriateness of test items prior to their implementation in operational settings. It typically involves piloting test items on a representative sample of test takers, as well as conducting qualitative and statistical evaluative analyses (see also Malec). Basic item analysis consists in computing and examining item facility (*IF*) and item discrimination (*ID*) values. According to the general guidelines for item evaluation in norm-referenced testing, ideal test items should have high item discrimination and moderate difficulty (see also, e.g., Haladyna). More specific procedures for evaluating MC questions involve determining the choice distribution (CD) patterns (as termed by Farhady). Roughly speaking, the correct answer of a well-performing MC item should mainly appeal to high-scoring test takers, whereas the distractors should be especially attractive to low-scoring test takers. Items where at least one of the options functions significantly differently are flagged for revision. The methods of evaluating MC options are discussed in more detail in Malec and Krzemińska-Adamek.

1. PURPOSE OF THE STUDY

The ultimate goal of the test development process was to compile three alternative test forms, each comprising 120 items. This paper focuses principally on pre-operational testing, which involved conducting statistical analyses of the data gathered from the pilot test forms. The specific purpose of the study was thus to select 360 best-performing test items on the basis of the results of item evaluation.

2. METHOD

2.1 Participants

The participants in the study were 2888 Grade 3 junior high school[1] students (aged 15) learning English within the *Youngster* project. They had

[1] The study reported in this paper was conducted before the latest reform of the educational system in Poland. The reform was aimed at, among others, the dissolution of junior high schools,

all had some previous experience of learning English as a foreign language, although their actual levels of proficiency varied. The participants came from rural areas characterised by low access to educational options, high unemployment and low income *per capita*. The above-mentioned factors are the main inclusion criteria used by the European Fund for the Development of Polish Villages to invite local councils to participate in the project.

2.2 MATERIALS

The content of the test developed for the *Youngster* project was, on the whole, intended to reflect the goals of foreign language learning outlined in the Polish National Curriculum, and, consequently, the requirements of the end-of-school exam. While, as has been explained above, the project is not aimed at preparing students for the exam as such, the requirements in question are important constituents of communicative competence, which is an overriding goal of foreign language instruction.

The test was designed to comprise three sections: vocabulary, grammar and language functions (also tested in the end-of-school exam). A variety of coursebooks used by teachers within the *Youngster* project served as the basis for the test content. The coursebooks in question represented four different levels of proficiency, corresponding roughly to the CEFR levels: A1, A2, B1, B2.

The initial pool of items comprised 1900 multiple-choice items, 910 of which were then selected in the course of qualitative analysis and compiled into 7 different test forms, each consisting of 130 test items targeting three language areas: grammar, vocabulary and functions. The pilot test forms included four groups of items of varying difficulty, referred to as A1, A2, B1, B2. The test items at different difficulty levels were not evenly distributed in the test forms, with items representing A2 and B1 levels constituting 30% of the items each, and A1 and B2 items constituting 25% and 15% respectively. Such a distribution of test items according to difficulty levels was justified on the grounds of the expected proficiency level of the test takers and the experience of the *Youngster* program leaders. The participants' anticipated level of proficiency was A2-B1. However, A1 and B2 levels were also included in the test with a view to tracking students whose proficiency level departed from the assumed range, and to allow teachers to react to their students' needs accordingly.

and reinstating 8-grade primary schools and 4-year secondary schools. At present, the *Youngster* project organises free English courses for students in the 8th grade of primary school.

2.3 Procedure

In order to ensure that the *Youngster* test was appropriate for the intended score uses and interpretations, qualitative and quantitative validity evidence of various types was collected and examined at every stage of its development. The specific stages preceding operational test use included drawing up the design specifications (which involved formulating the item-writing guidelines), producing the test items (which involved item-writer training, item writing, obtaining expert judgements of item quality, redrafting the items, selecting the best items for pilot testing), and pre-operational testing.

2.3.1 *Compiling the pilot test forms and qualitative analysis*

The initial stage of developing the test involved the formulation of item-writing guidelines. The guidelines in question specified the types of stems, and types and number of options that were to be found in the test. Additionally, they informed the test writers about some potential challenges or areas of difficulty (e.g. pertaining to designing good quality distractors). After the initial versions of the test items were written, they were critically evaluated by three independent experts. As regards the roles of the experts who reviewed the drafts, they were twofold: (1) to judge the degree to which the items represented the target domains, i.e. the material from the coursebooks and (2) to assess critically the formulation of the test items (length of options, grammatical consistency of options, absence of double keys, avoidance of stereotyped or emotionally loaded responses, and the like – see also, e.g., Allan; Coombe, Folse and Hubley for a further discussion on common MC item violations). In the next step, the initial versions of the test items were redrafted according to the critical comments made by the experts, with a view to ensuring their best possible quality. Finally, in the course of the qualitative analysis outlined above, 910 best items were selected to be piloted.

2.3.2 *Administration of pilot test forms*

The pilot test was administered online under the supervision of teachers participating in the project. Every test taker received one of seven test forms, each consisting of 130 items. Every test form was taken by a group of no fewer than 400 students. The time limit for completing the test was

60 minutes. The test takers did not have the possibility to revise their answers once they had been provided.

2.3.3 *Scoring and statistical analysis*

It should be pointed out that the present study employed statistical methods which are appropriate to norm-referenced measurement because the test was first to be utilised for placement purposes (cf. Brown and Hudson 251). However, since the content of the test was actually based on the *Youngster* project teaching materials, it was legitimate to make use of the test scores also for criterion-referenced interpretations (as a measure of progress in learning).

The pilot test items were scored 0-1, where 0 meant incorrect or no answer. The data collected in the course of test administration was additionally scrutinised for cases where participants exhibited a tendency to choose the same answers throughout the test, apparently displaying a negative attitude to the test-taking event. When the number of identical answers (A, B, or C) from a single test taker exceeded 80% of all the responses, the scores obtained from this participant were excluded from further analysis. The overall number of such cases was 79, ranging from 5 in Group 5 to 16 in Group 2.

The main aim of the study was to select items of the best quality in terms of option performance. However, it was also important to verify item difficulty at each level of proficiency. More specifically, items at a given level were expected to have similar *IF* values across the seven test forms. On the other hand, within each test form, there had to be significant differences between *IF* values across the four levels of difficulty. These analyses were performed in SPSS, with items as cases and *IF* as the dependent variable. Depending on whether or not the data deviated significantly from normality, the differences were analysed by means of either parametric (ANOVA) or non-parametric (Kruskal-Wallis) tests.

Option performance was assessed with the aid of Excel using several methods of classical item analysis (see also Malec and Krzemińska-Adamek). The first method was the point-biserial correlation calculated for the key (PB_C) and for each of the distractors (PB_{DC}). This method was employed to obtain statistical measures of item and distractor discrimination. In the case of the distractors, the calculations were based on a modified formula proposed by Attali and Fraenkel. In addition to the point-biserial correlation, the frequencies of each option being selected in five score groups (established on the basis of total test scores) were analysed statistically. The aim of this procedure was to see whether the frequencies increased or decreased significantly from lower to higher score groups. This method of evaluating option performance relies on the chi-square test and is often referred to as the categorical analysis of the trace line (Haladyna and Rodriguez) The final statistical method used in this study was the analysis of choice means. A choice mean is the average total score obtained for the test takers who selected a particular option. A well-constructed MC item should have a higher choice mean for the correct answer than for any of the distractors. One way of determining the statistical significance of the differences between all the choice means is through the Pearson correlation between total scores and choice means substituted for option choices (Haladyna; see also Malec on conducting this procedure in Excel).

All the statistical methods produced numerical estimates which helped to determine whether the analysed test items were eligible for inclusion in the final version of the test. The specific selection criteria were defined as in Table 1.

Table 1: Item selection criteria

Criterion	**Description**
(1) $PB_C \geq .30$	Item discrimination (point-biserial correlation) should be at least .30 (cf. Ebel; Niemierko).
(2) $PB_{DC} < -.10$	The point-biserial correlation should be negative for each distractor, preferably below −.10.
(3) χ^2_{FREQ}*	The chi-square test for the frequencies of each option in five score groups should be statistically significant at .05.
(4) $CM_{KEY} - CM_{DIS} > 5$ & $r_{CM\ TS}$*	The choice mean of the key should be higher than the choice means of the distractors by at least 5 points AND the correlation between choice means (substituted for option choices) and total scores should be statistically significant at .05.
(5) $DIS_{LOW} \geq 10\%$	Distractors must be selected by at least 10 percent of low-scoring test takers.
(6) $DIS_{HIGH} < 33\frac{1}{3}\%$	Distractors must be selected by fewer than one third of high-scoring test takers.

Note: The asterisk (*) is used to indicate statistical significance at an alpha level of .05.

As can be seen in Table 1, there were two additional criteria. Given that statistical analyses may not be able to indicate that an option is selected too infrequently, it was necessary to set arbitrary standards in this regard (cf. Haladyna and Rodriguez 355). First, in view of the fact that distractors should appeal especially to weaker students, only those items were accepted whose distractors had been selected by at least 10 percent of students in the lower group (Criterion 5). Second, distractors should not attract too many test takers in the upper group. Therefore, items with a distractor selected by at least one third of high scorers were rejected (Criterion 6).

It bears pointing out that in this particular test development setting, priority was given to picking out items of the best quality rather than to modifying those which malfunctioned. The reason for this was that there was no possibility to organise additional pilot test sessions.

3. RESULTS AND DISCUSSION

3.1 Test statistics

The basic test statistics were calculated for each language area and for each level of proficiency. The same pattern of difficulty was found in every test form: vocabulary items were the easiest, whereas grammar items were the most difficult. Form 1 had the highest overall mean *IF* (.54), whereas Form 4 had the lowest (.48). Form 4 also had the highest omission rate. This may suggest that either Form 4 was the most difficult of all or the students who completed it represented a marginally lower level of proficiency. The significance of these differences is tested in Section 3.2 below.

As can be seen in Table 2, the mean difficulty of test items increased with the level of proficiency, although in some cases these differences were not particularly high (e.g. B1 and B2 in Form 2). Another observation that can be made is that B1 and (especially) B2 items may have been too difficult for these test takers, as evidenced by relatively high omission rates. This suggests that the final version of the test should include only a small number of B2 items.

Table 2: Descriptive statistics for levels of proficiency

Test Form	***N***	**Level**	***k***	**Mean *IF***	***OR***	***SD***
F1	403	A1	32	.73	1.05	5.33
		A2	39	.58	4.84	8.53
		B1	39	.46	9.43	8.87
		B2	20	.39	12.38	3.98
F2	397	A1	32	.68	0.92	6.26
		A2	39	.55	4.33	8.99
		B1	39	.42	9.08	7.34
		B2	20	.40	12.64	4.35
F3	401	A1	32	.67	1.11	6.01
		A2	39	.54	4.92	8.10
		B1	39	.46	9.92	8.21
		B2	20	.37	13.74	4.13
F4	404	A1	32	.64	0.90	5.59
		A2	39	.51	5.51	8.08
		B1	39	.40	12.15	7.93
		B2	20	.35	15.85	4.11
F5	402	A1	32	.67	1.21	5.80
		A2	39	.52	4.33	7.79
		B1	39	.41	10.91	8.02
		B2	20	.35	13.93	3.83
F6	405	A1	32	.69	0.68	5.86
		A2	39	.55	3.13	7.96
		B1	39	.46	9.91	8.28
		B2	20	.37	14.67	3.99
F7	397	A1	32	.68	0.56	5.94
		A2	39	.55	3.73	7.60
		B1	39	.43	9.67	7.67
		B2	20	.38	13.10	3.89

N = number of students; *k* = number of items; OR = omission rate; SD = standard deviation

Cronbach's alpha was used to estimate the internal consistency of the scores. When calculated for an entire language area (including all levels of proficiency) in any given test form, the coefficient was always higher than .80. However, reliability was generally lower for smaller sections of the test (language area by level of proficiency), particularly in the case of grammar and B2. One of the reasons for this is that B2 sections had only a few items. Moreover, since grammar and B2 were the most difficult parts of the tests, guessing may have contributed to the low reliability estimates (e.g. Wright). Since adding more items to sections that were already very difficult would have defeated the purpose of increasing reliability, the only way to reduce error variance was by removing those items whose options were not statistically functional.

3.2 Differences between test forms and levels of proficiency

The differences between test forms and levels of proficiency were tested by (1) comparing *IF* values in different test forms at each specific level of proficiency, and (2) comparing *IF* values at different levels of proficiency within each specific test form. The mean differences in question are given in Table 2 above, under 'Mean *IF*': vertically for each level of proficiency and diagonally for each test form. The Shapiro-Wilk test was used to detect deviations from normality because it is more powerful than the Kolmogorov-Smirnov test (Field 148). The results of this test for each level of proficiency in each test form are given in Table 3. The last row and column show the statistical procedures (with results) that were appropriate to the data given: the Kruskal-Wallis test where (at least some of) the data deviated from normality and ANOVA elsewhere.

Table 3: The Shapiro-Wilk test results and statistical procedures (with results) used to compare differences between test forms and levels of proficiency

	A1	**A2**	**B1**	**B2**	**Procedure and result**
Form1	.001	.701	.349	.949	$H(3) = 54.818$, $p < .001$
Form2	.364	.166	.431	.635	$F(3) = 33.227$, $p < .001$
Form3	.774	.574	.558	.428	$F(3) = 28.168$, $p < .001$

Form4	.044	.768	.625	.270	$H(3) = 40.598$, $p < .001$
Form5	.029	.425	.448	.197	$H(3) = 48.123$, $p < .001$
Form6	.016	.287	.424	.822	$H(3) = 46.524$, $p < .001$
Form7	.004	.438	.184	.997	$H(3) = 43.464$, $p < .001$
Procedure and result	$H(6) = 6.057$, $p = .417$	$F(6) = 1.379$, $p = .223$	$F(6) = 1.566$, $p = .157$	$F(6) = 0.802$, $p = .570$	

H = the Kruskal-Wallis test; *F* = ANOVA

The results given in Table 3 are as expected: there were significant differences between levels of proficiency and no significant differences between test forms. However, *post hoc* tests (with a Bonferroni correction) revealed that there were no significant differences between the following levels of proficiency: A1-A2 (Form 4 and 7), A2-B1 (Form 6), and B1-B2 (every test form). To avoid this, items whose *IF* values were close to the mean *IF* at a different level of proficiency were rejected. In addition, changes were made to the proficiency classification of some of the items. In particular, a number of the original B1 items were reclassified as B2 on the grounds of their low *IF* values. This decision may be justified inasmuch as the levels of proficiency represented by the coursebooks from which the target language material was sampled are to a large extent arbitrarily defined. In addition to that, there is no certainty as to whether a particular book contains solely language items belonging to the level specified on the cover.

3.3 Selecting the final test items

IF values were compared to *IF* grand means at each level of proficiency (for the entire set of data), which were .68 (A1), .54 (A2), .43 (B1), and .37 (B2). As mentioned above, on the basis of their (comparably low) *IF* values, some items had their proficiency classification changed. More particularly, two items were shifted from A2 to B1 and eight items from B1 to B2.

Next, the selection criteria given in Table 1 were applied in six stages to all the 910 items. Table 4 presents the results.

Table 4: The application of the selection criteria to 910 test items

Stage	Criterion	Initial items	Unsatisfactory items (overall)	Items rejected at stage	Remaining items
1	$PB_C \geq .30$	910	268	268	642
2	$PB_{DC} < -.10$	642	133	6	636
3	χ^2_{FREQ}*	636	526	286	350
4	$CM_{KEY} - CM_{DIS} > 5$ & $r_{CM\,TS}$*	350	139	0	350
5	$DIS_{LOW} \geq 10\%$	350	51	10	340
6	$DIS_{HIGH} < 33\frac{1}{3}\%$	340	141	4	336

Note: The asterisk (*) signifies statistical significance at an alpha level of .05.

Specifically, Table 4 shows the number of all items available at each stage (Column 3), number of items of the entire set of 910 items which did not meet the criterion (Column 4), number of items rejected at the given stage that were not rejected at previous stages (Column 5), and number of the remaining items (Column 6). Only 336 items fulfilled all of the criteria, which was not acceptable because that meant insufficient items for several sections of the final test. Therefore, the decision was made to remove Criterion 3, which was very stringent: there was an inordinately large number of items (i.e. 248) which failed to fulfil this criterion alone, even though they fulfilled all the remaining criteria (see also Malec and Krzemińska--Adamek for more details).

In order to compensate (at least partly) for the removal of the chi-square test as a selection criterion, the threshold for PB_{DC} was lowered to −.15. This was supposed to ensure acceptable distractor discrimination. The results of the final selection of items are given in Table 5.

Table 5: The application of the final selection criteria to 910 test items

Stage	Criterion	Initial items	Unsatisfactory items (overall)	Items rejected at stage	Remaining items
1	$PB_C \geq .30$	910	268	268	642
2	$PB_{DC} < -.15$	642	209	33	609
3	$CM_{KEY} - CM_{DIS} > 5$ & $r_{CM\,TS}$*	609	139	0	609
4	$DIS_{LOW} \geq 10\%$	609	51	25	584
5	$DIS_{HIGH} < 33\frac{1}{3}\%$	584	141	19	565

Note: The asterisk (*) signifies statistical significance at an alpha level of .05.

The application of this modified set of criteria yielded enough items for each section of the final test. Of the 565 items which met all the criteria, the final 360 included those items whose *IF* values were closest to the *IF* mean at each level.

The purpose of the last statistical test was to compare the differences between the difficulties of the levels of proficiency in the final set of 360 items. The relevant means and standard deviations are given in Table 6.

Table 6: Descriptive statistics for the levels of proficiency in the final set of items

	Number of items	**Mean *IF***	**SD**
A1	69	.68	.079
A2	120	.56	.072
B1	117	.48	.064
B2	54	.39	.070
Total	360	.53	.116

The Shapiro-Wilk test was not significant for any of the levels of proficiency, hence a parametric ANOVA could be run. Its result, $F(3) = 200.56$, $p < .001$, as well as all *post hoc* tests with a Bonferoni correction, confirmed that there were significant differences between the four levels of proficiency in the final set of 360 items.

CONCLUSION

When a pool of the best-performing items had been established, three operational test forms were created, each consisting of 120 questions. These test forms were compiled by the project organisers through a random selection of items from each specific language area and level of proficiency. Each participant of the programme was to take all the three test forms at three different points in time: the beginning of the school year, the end of the first semester and the end of the school year. Thus, the new instrument was intended to perform both placement and progress functions within the project.

The new test has already been successfully implemented. It informs the teachers about the current proficiency level of the students who begin learning English within the project, which enables an adequate choice of materials and response to individual instructional needs. Thus, the instrument can be tentatively said to have good predictive validity, in the sense that its results find a confirmation in everyday classroom work. In addition, the test has been found to accurately measure the progress that the students have made in the course of learning, which has been expected by the organisers of the *Youngster* project. All in all, the test has proved to be a good-quality instrument serving a number of purposes. Achieving these goals, however, was only possible thanks to a scrupulous process of planning, designing, reviewing, and finally – statistically validating the test items. Some of the statistical procedures described in this paper, particularly those which are relevant to evaluating the performance of multiple-choice options, were carefully adjusted to the specific testing situation.

BIBLIOGRAPHY

Allan, Alastair. "Development and Validation of a Scale to Measure Test-Wiseness in Efl/Esl Reading Test-Takers." *Language Testing*, vol. 9, no. 2, 1992, pp. 101–22.

Attali, Yigal, and Tamar Fraenkel. "The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an Alternative." *Journal of Educational Measurement*, vol. 37, no. 1, 2000, pp. 77–86.

Brown, James Dean, and Thom Hudson. *Criterion-Referenced Language Testing*. Cambridge University Press, 2002.

Coombe, Christine, Keith Folse, and Nancy Hubley. *A Practical Guide to Assessing English Language Learners*. University of Michigan Press, 2007.

Davison, Mark L., Gina Biancarosa, Sarah E. Carlson, Ben Seipel, and Bowen Liu. "Preliminary Findings on the Computer-Administered Multiple-Choice Online Causal Comprehension Assessment, a Diagnostic Reading Comprehension Test." *Assessment for Effective Intervention*, vol. 43, no. 3, 2018, pp. 169-81.

Ebel, Robert L. "Procedures for the Analysis of Classroom Tests." *Educational and Psychological Measurement*, vol. 14, 1954, pp. 352–64.

Farhady, Hossein. "Principles of Language Assessment." *The Cambridge Guide to Second Language Assessment*, edited by Christine Coombe, Peter Davidson, Barry O'Sullivan, Stephen Stoynoff, Cambridge University Press, 2012, pp. 37-46.

Field, Andy. *Discovering Statistics Using Spss*. 3rd ed, Sage Publications Ltd, 2009.

Frizelle, Pauline, Clodagh O'Neill, and Dorothy V.M. Bishop. "Assessing Understanding of Relative Clauses: A Comparison of Multiple-Choice Comprehension Versus Sentence Repetition." *Journal of Child Language*, vol. 44, no. 6, 2017, pp. 1435-57.

Gyllstad, Henrik, Laura Vilkaitė, and Norbert Schmitt. "Assessing Vocabulary Size through Multiple-Choice Formats: Issues with Guessing and Sampling Rates." *ITL – International Journal of Applied Linguistics*, vol. 166, no. 2, 2015, pp. 278-306.

Haladyna, Thomas M. *Developing and Validating Multiple-Choice Test Items*. 3rd ed, Lawrence Erlbaum, 2004.

Haladyna, Thomas M., and Michael C. Rodriguez. *Developing and Validating Test Items*. Routledge, 2013.

Kenyon, Dorry M., and David MacGregor. "Pre-Operational Testing." *The Routledge Handbook of Language Testing*, edited by Glenn Fulcher and Fred Davidson, Routledge, 2012, pp. 295-306.

Koyama, Dennis, Angela Sun, and Gary J. Ockey. "The Effects of Item Preview on Video-Based Multiple-Choice Listening Assessments." *Language Learning & Technology*, vol. 20, no. 1, 2016, pp. 148-65.

Malec, Wojciech. *Developing Web-Based Language Tests*. Wydawnictwo KUL, 2018.

Malec, Wojciech, and Małgorzata Krzemińska-Adamek. "A Practical Comparison of Selected Methods of Evaluating Multiple-Choice Options through Classical Item Analysis." *Practical Assessment, Research, and Evaluation*, vol. 25, no. 1, Art. 7, 2020, pp. 1-14.

Niemierko, Bolesław. *Pomiar Wyników Kształcenia* [Measurement of Learning Outcomes]. Wydawnictwa Szkolne i Pedagogiczne, 1999.

Sonbul, Suhad, and Norbert Schmitt. "Explicit and Implicit Lexical Knowledge: Acquisition of Collocations under Different Input Conditions." *Language Learning*, vol. 63, no. 1, 2013, pp. 121-59.

Wright, Robert J. *Educational Assessment: Tests and Measurements in the Age of Accountability*. Sage Publications, Inc., 2007.

TWORZENIE TESTU Z JĘZYKA ANGIELSKIEGO DLA UCZESTNIKÓW PROGRAMU *YOUNGSTER* W POLSCE

Streszczenie

Artykuł opisuje proces tworzenia testu z języka angielskiego jako obcego, przeprowadzanego za pośrednictwem internetu, przeznaczonego dla uczestników programu *Youngster* w Polsce. Program powstał z myślą o zapewnieniu bezpłatnej nauki języka angielskiego dla nastoletnich uczniów zamieszkujących tereny wiejskie. Jednym z wymogów programu jest zagwarantowanie skutecznych metod i instrumentów oceniania, regularnie wykorzystywanych przez nauczycieli na trzech różnych etapach nauki – na początku roku szkolnego, na końcu pierwszego semestru i na końcu roku szkolnego. Celem badania opisanego w niniejszym artykule było opracowanie testu, który mógłby pełnić nie tylko funkcje plasujące, ale także służyć skutecznej ocenie postępów uczniów biorących udział w programie. Artykuł omawia procedury związane z testowaniem próbnym, wykonanym w celu przeprowadzenia analizy jakości poszczególnych zadań, i koncentruje się przede wszystkim na tworzeniu zadań typu wybór wielokrotny.

Słowa kluczowe: tworzenie testów językowych; zadania wyboru wielokrotnego; analiza zadań; ocena jakości dystraktorów.

THE DEVELOPMENT OF AN EFL TEST FOR THE *YOUNGSTER* PROJECT IN POLAND

Summary

The article is a report on the development of an online assessment instrument, namely a test of English as a foreign language intended for the participants of the *Youngster* project in Poland. The project has been launched with a view to providing free English education for learners in rural areas. One of the requirements of the project is the provision of effective assessment procedures, regularly conducted at three different points in time – the beginning of the school year, the end of the first semester, and the end of the school year. The purpose of the study was therefore to develop an instrument which could serve both placement and progress functions within the project. This article describes the activities performed prior to the test's operational administration, and focuses primarily on the development of multiple-choice items.

Keywords: EFL test development; multiple-choice format; item analysis; distractor evaluation.