The use of the Iteman program in designing a General Foreign Language placement test

Monika Kusiak-Pisowacka

The Institute of English Studies, Jagiellonian University Kraków, Poland

Abstract- The paper is a report on the project which involved the process of designing and piloting a General Competence Test intended to serve as a placement test in teaching Polish intermediate foreign language learners. All the actions taken to design and pilot the test are discussed in detail. The use of Iteman 4, a program designed to do a detailed item test analysis, is described. Problems encountered by the teacher during the process of designing, piloting and preparing the final version of the test are discussed. It is hoped that this paper will elucidate the advantages of tests constructed by foreign language teachers themselves and will encourage them to produce their own tests tailored for their teaching needs.

Index Terms- foreign language evaluation, Iteman program, piloting tests, placement tests

I. Introduction

This article is about evaluation, which is a crucial element of foreign language (FL) education, particularly the one organized within formal education. Tests are the most common, although not the only forms of measuring learners' FL competence. Both internal, administered by teachers, and external, high stakes tests administered by the outside institution, e.g. the Ministry of Education constitute an important part of teachers' and learners' school life. My own experience with language testing has revolved around a number of individuals involved in school education, most of them having to cope with testing as an omnipresent component of language teaching programs. As a teacher trainer I have met many pre-service and in-service teachers who experienced difficulty in designing tests for their students; as a supervisor of BA and MA seminars I have advised students who need to develop tests for their teaching and research purposes. Most of the teacher trainees and experienced educators that I have met seem to lack knowledge and skills necessary to design their own FL tests as well as confidence to evaluate critically ready tests available in published FL materials. In the next sections of this paper I describe the actions I took to design a general competence test for my own teaching purposes. In the introduction to this report, I look at basic issues related to testing, namely qualities of tests and the role that tests play in school education.

A. Qualities of tests

The most important qualities of tests discussed in the literature, e.g. by Bachman and Palmer (1996), and Alderson, Clapham and Wall (1995), are: reliability, validity, authenticity, interactiveness, impact and practicality. For the purpose of the study presented in this paper, reliability, validity, authenticity, impact and practicality will be discussed.

"Reliability is often defined as consistency of measurement" (Bachman and Palmer 1996: 19). Validity is the extent to which a test measures what it is intended to measure. A test construct refers to an ability that the test intends to measure. Thus, construct validity is defined as "the extent to which we can interpret a given test score as an indicator of the ability(ies), or construct(s), we want to measure" (Bachman and Palmer 1996: 21).

As regards authenticity, it refers to the extent to which the test resembles the use of target language in real life situations, i.e. outside the test itself. Another quality of tests important in designing tests is their impact on those involved in testing situations, i.e. test takers and teachers (a micro level of impact) as well as the educational system within which tests are administered and society (a macro level of impact). The last test quality to be discussed in this paper is practicality. When assessing practicality, we have to consider resources (human resources, material resources and time) needed to develop and administer the test. A test is practical if the resources required do not exceed the available resources.

In this short discussion concerning test features, it is crucial to add that test qualities are interrelated. Bachman (1990: 289) says: "While validity is the most important quality of test use, reliability is a necessary condition for validity, in the sense that test scores that are not reliable cannot provide a basis for valid interpretation and use." Similarly, Hughes (1989) claims that if the test is not reliable, it cannot be valid. The discussion shows that testing as a domain of language learning and teaching is a complex phenomenon.

B. The role of tests in formal education

In foreign language evaluation tests can be applied for different purposes. Below the following types of tests are discussed: placement, progress, achievement, proficiency and diagnostic. The discussion is based on Alderson, Clapham and Wall (1995). Placement tests aim to assess students' level of language competence so that learners can be placed in appropriate language groups. They are popular in language centers, which need to check their students' competence in order to place them

in appropriate level groups. At schools placement tests are usually administered at the start of a school year. The most popular tests administered in school education are progress and achievement tests. Progress tests are given at various stages throughout a language course to see what and how much the learners have learnt. Achievement tests serve the same function, but are usually given at the end of the course. The content of both tests is based on the material covered during the course and/or the textbook. The aim of proficiency tests is to assess whether the students have reached a given level of FL competence and how well they can function in certain areas which require the use of a foreign language. To conclude, "achievement assessment is oriented to the course. It represents an internal perspective. Proficiency assessment on the other hand is assessment of what someone can do/knows in relation to the application of the subject in the real world. It represents an external perspective" (*Common European Framework of Reference* 2001: 183). In school conditions, foreign language teachers are more interested in achievement assessment because it is the type of test that can provide them with feedback about their teaching. Proficiency tests are administered in the form of external exams, the example of which are the junior secondary school leaving exam and the senior secondary school leaving exam, called in Polish "matura". The last type of test to be discussed in this section is a diagnostic test. Its aim is to identify those areas in which students need help. Since constructing this type of test requires specialized skills on the part of the teacher, very often achievement and proficiency tests are used for this purpose.

C. The use of Iteman in preparing tests

ITEMAN 4.3.0.3 was originally created by Assessment Systems Corporation. The most popular versions of the tool are 4.3, 4.2 and 4.1. It is a software program designed to provide detailed item and test analysis reports using classical test theory (CTT), the information which can be helpful in the evaluation of the quality of test items, and tests as a whole. The program examines the tests' psychometric characteristics. It also produces summary output regarding the examinee scores, including reliability analysis, analysis of domains (content areas), and frequency distributions. An undeniable advantage is that it can be downloaded from the Internet for free. It allows one to save the reports in RTF, which enables test constructors to prepare a comprehensive report to stakeholders, such as head teachers or external experts.

Iteman has been used in a number of exam projects, mostly those which involved the construction and evaluation of high-stakes language exams. For example, it was used in Poland during the standardization of the final practical English examination in the Kraków cluster of colleges. The whole process has been described in detail in Defty and Kusiak (1997), and Kusiak and Jurek-Kwiatkowska (1998). Unfortunately, the Iteman program is still a novelty for foreign language teachers, as demonstrated by Aulia, Sukirlan and Sudirman (2014) in their analysis of the quality of teacher-made reading comprehension tests. On the basis of their study the authors conclude that it is important that teachers of English should be trained to use the Iteman program since this ability can improve the quality of the tests they use and consequently the quality of their teaching. This argument seems true also in relation to Polish teachers of English. During my work as a teacher trainer I have not encountered a teacher who would be able to use Iteman and interpret the analyses produced by the program, although quite a number of teachers enthusiastically use various computer software programs in their teaching.

II. THE DESCRIPTION OF THE PROJECT

The aim of the project was to construct a General Language Competence Test which could allow me to assess the level of general proficiency of English of a group of intermediate students and within this group distinguish several levels of proficiency. The test falls within the category of placement tests. Thus, the main aim of the project was to construct the test and to check the following issues related to the quality of the test:

1. the capacity of the test to distinguish different English proficiency groups,
2. the reliability of the test items,
3. the administration conditions, e.g. timing, instructions, etc.,
4. the marking procedure and the answer key.

This section will present the blueprint I followed in the construction of the test. It will state the purpose of the test, the description of subjects and the description of the test content.

A. Test specifications

The main purpose of this written test was to asses students' level of language proficiency, so that the whole subject sample could be divided into three groups according to their language proficiency. To ensure this, the test was supposed to discriminate well and have a wide spread of scores.

The test was intended for secondary school students studying English as an obligatory subject from four to six lessons a week at intermediate, B1 level of English proficiency (according to *Common European Framework of Reference* 2001). The test was intended to assess the learners' knowledge of vocabulary, grammar, syntax as well as the sensitivity to discourse cohesion. To ensure reliability, it was decided that the test items would include a variety of item types, i.e. multiple choice questions, gapped texts and cloze tests. Most of the texts used in the test would be drawn from authentic materials or would adapt such materials. It was attempted to choose texts which would be of general interest, and would not require specialist knowledge and/or vocabulary. It was decided that all the test items could be objectively marked and a clear answer key would be produced. Special attention was paid to writing clear instructions and providing examples of items where necessary. Time allowed for the test was planned not to exceed forty five minutes.

B. The first draft of the test

The first draft of the test consisted of five tasks:

1. multiple choice questions testing understanding vocabulary in context,
2. a cloze test assessing the ability to use grammar tenses,
3. two gapped texts with multiple choice answers testing syntax, coherence and vocabulary,
4. a cloze test testing the knowledge of vocabulary, grammar and sensitivity to text coherence.

Altogether the test had 48 items; the number was considered sufficient to obtain valid test results. It was assumed that the test would take 45 minutes. A complete version of the first draft of the test is provided in the Appendix.

C. Subjects and procedure

The test was trialed on 21 students from an EFL school. The students constituted a relatively homogenous group at the intermediate level of English proficiency. The choice of a language course group rather than a regular school class was determined by the main purpose of the test, i.e. obtaining a wide spread of scores. A more homogenous group of subjects seemed to provide better conditions to test the discriminating capacity of the test. A language course group was regarded more homogenous than a typical secondary school class, which is often of a mixed language ability.

The test was administered by the teacher conducting the language course. In order to ensure that the students should take the test seriously, the students were informed about the test in advance and the test was introduced as a mid-term progress proficiency test.

III. RESULTS OF THE ANALYSIS

All the marking was done by the author of this paper according to the answer key prepared beforehand (see Appendix). The time allocated to the test, i.e. 45 minutes, proved sufficient for completing the test. Most of the subjects managed to finish the test even after 35 minutes. Time allocated for a test is an important trailing condition. The situation in which the students are not given enough time may result in a high proportion of unattempted answers and consequently inflate reliability indices (as explained by Crocker and Algina 1986). The testing conditions allowed me to eliminate this unfavourable possibility.

The interview with the teacher who administered the test revealed the flaw in the instruction of Exercise 2. Despite the example provided, the students did not find the instructions clear enough, i.e. they were not sure how many words they were supposed to fill in each gap. It was decided that the instruction would be developed. It would state that it is possible to use more than one word, if the form requires it.

The result data were analyzed using the Microcat computer program ITEMAN. Below the descriptive statistics results are presented (see Table 1). They provide the information about the score range of the test scores and the reliability of the test.

Table 1. The distribution statistics for the students taking the General Language Competence Test (the first version).

No of subjects	21
No of items	48
Mean Raw Score	25.62
Raw Score S.D.	5.85
Mean Score as a %	53.44 %
Percentage Score S.D.	12.19
Median Raw Score	24.00
Minimum Score	14
Minimum Score as a %	29 %
Maximum Score	35
Maximum Score as a %	73
Alpha	0.77

Table 2. The ITEMAN analysis of the chosen items of the results of the General Language Competence Test (the first version).

Item Statistics				Alternative Statistics					
Seq. No.	Scale -Item	Pcnt Correct	Disc. Index	Point Biser.	Alt.	Pcnt Total	Endorsing Low	Point High	Biser. Key
6	0-6	86	-.17	-.17	A	86	100	83	-.17 *
					B	0	0	0	
	CHECK THE KEY				C	10	0	17	.10
	a was specified, d works better				D	5	0	0	.13 ?
					Other	0	0	0	
30	0-30	0	.00		A	43	17	17	-.08
					B	0	0	0	
					C	52	83	67	-.08
					D	0	0	0	*
					Other	5	0	0	.36
34	0-34	24	-.17	-.04	A	67	67	67	-.03
					B	24	33	17	-.04 *
	CHECK THE KEY				C	0	0	0	
	b was specified, d works better				D	5	0	17	.21 ?
					Other	5	0	0	-.06

The mean, i.e. 53.44 %, show that the test is of a suitable level of difficulty. The standard deviation and ranges (i.e. the difference between the maximum and minimum scores) indicate that scores range from 29% to 73%, which seems satisfactorily wide. The histogram (see Figure 1) also demonstrates that the scores are spread evenly; they are not clustered together at the top of the distribution (a negative skew) or at the bottom of the distribution (a positive skew). Such an even distribution is appropriate for a language competence test which is intended to identify students at different levels of proficiency. This finding indicated that the test fulfilled its basic purpose, i.e. a wide enough spread of scores to distinguish three different proficiency levels among the students.

Test item analysis showed that the test items worked well. Twenty nine items reached satisfactorily high level of mean point biserial correlations, ranging from .21 to .67. Eleven items had relatively low mean point biserial correlations, e.g. item no 48, ranging from .01 to .18. Five items (items no 6, 13, 25, 34, 35) had negative discriminations, i.e. more low group students (those who perform worst on the test) were correct than top group students (those who perform best on the test). Negative discriminations indicated that there was something wrong with these items. Three items (23, 30, 31) did not discriminate at all; i.e. their discrimination index equals 0. Item 23 was too easy and was answered correctly by all the students, and the other two, i.e.

items 30 and 31, were too difficult and nobody chose the expected correct answers. Table 2 presents the item analysis of items no 6, 30 and 34.

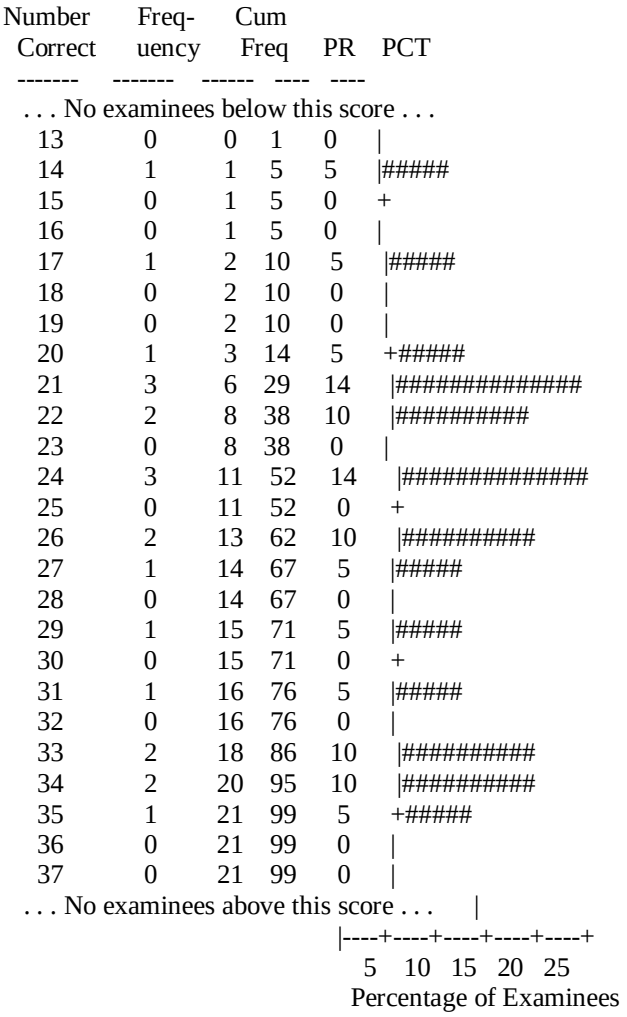

Figure 1: The Score Distribution Table for the students taking the General Language Competence Test (the first version).

The ITEMAN analysis results enabled me to revise the first draft of the test. The items with very low or negative discrimination indices were changed or discarded. Multiple choice questions were examined for the students' performance on the items' distractors. When the distractor did not attract any answers, they were revised in order to improve low discrimination indices. For example, item 34 (see Table 2) discriminate negatively with a point biserial of -.04. No one chose answer C. Answer B, which is the correct one (indicated by an asterisk) attracted more low group students than top group students, which consequently gave this distractor a negative discrimination index, i.e.-.04. The results also indicated that although distractor B was specified as the correct answer, distractor D turned out to work better; and it had a positive discrimination index, i.e. .21, which is not expected from a distractor which is not the correct answer.

The reliability of the test was measured by means of the Kuder Richardson (20) reliability index , in Table 1 indicated as the alpha index. The reliability index, which is related to the discrimination indices, seemed satisfactory, i.e. .77. This result seemed true, especially when I took into consideration the fact that the students were allowed as much time as they needed to complete the test (see my explanation of the relation of this factor and test reliability at the beginning of this subchapter).

To sum up, the pretesting of the General Language Competence proved that the test was a good placement test that would enable me to divide the students into several different proficiency levels. The reliability and discrimination indices provided valuable information about the general quality of the test and the students' performance on particular items. This led to the revision of the test and the answer key, the process which is not described in this paper

APPENDIX

Appendix: The General English Competence Test with the Answer Key (the first version)

Exercise 1. In questions 1-10 each sentence has an underlined word. Below each sentence there are four other words marked A, B, C and D. Choose the one word that best keeps the meaning of the original sentence if it is substituted for the underlined word. Mark your answer in a clear way.

An example:
We did not expect such an abrupt rise in food prices.
a. important; b. serious; c. **sudden;** d. considerable

1. Beekeeping has become a sophisticated operation. It requires special skill and a lot of equipment.
   a/ expensive; b/ complex; c/ scientific; d/ profitable
2. Smallpox has been universally eradicated.
   a/ eliminated; b/ pushed over; c/ assimilated; d/ verified
3. For centuries people have exploited the ability of certain herbs to improve stamina.
   a/ searched for; b/ taken advantage of; c/ improved; d/ argued for
4. Natural sponges are considered indispensable for cleaning certain scientific instruments.
   a/ impossible; b/ difficult; c/ essential; d/ incredible
5. Standard IQ tests have been denounced by many educators as being culturally biased.
   a/ encouraged; b/ condemned; c/ exemplified; d/ claimed
6. The question of when humans first inhabited the North American continent is intriguing.
   a/ fascinating; b/ invigorating; c/ entertaining; d/ improbable
7. Large carnivorous aquatic creatures have been seen in Loch Ness since the Middle Ages.
   a/ acrobatic; b/ muscular; c/ marine; d/ ancient
8. They were spotted by the police as they were leaving the bank.
   a/ imprisoned; b/ caught; c/ noticed; d/ photographed
9. If you think she is doing it because she loves you, you are deceiving yourself.
   a/ hurting; b/ cheating; c/ flattering; d/ criticizing
10. The need for adequate housing is very acute in areas where catastrophes have occurred.
    a/ faults; b/ disasters; c/ dangers; d/ tornadoes
11. Pagan, the ancient capital of Burma, was widely renowned for its 5,000 Buddhist temples.
    a/ discussed; b/ worshipped; c/ acclaimed; d/ visited

Exercise 2. Read the text and complete the gaps with the right forms of the words provided in the brackets. The first has been done for you.

The big clock struck five. The hall (pack) …***was packed***…. with people who 12/ (come) ……. to listen to the concert. Carl 13/ (take) ……. his seat on stage. He was a pianist, but he really wished he 14/ (play) ……. the violin. When the performance was over, the people shouted "Bravo". The audience wanted him to play another piece, but while they 15/(shout) ……. "Encore" Carl put on his coat. The 16/ (cheer) ……. did not encourage him to play more. Outside the hall a young admirer asked him for his autograph. She also said that se 17/ (wish) ……. she 18/ (have) ……. his talent to play the piano. Carl smiled. He felt better about 19/ (be) ……. a pianist.

Exercise 3. Read the text and decide which answer A, B, C or D best fits each space. Circle your answers.

Computers in sports
More and more athletes and their coaches are using computers to help in training for sports. 20/ ……., almost every major league baseball and football team uses computers. So 21/ ……. many well-known golf and tennis players. Olympic athletes and coaches use computers too. Computers 22/ ……. the place of people as coaches, but they can help in many different ways. The computer is a(n) 23/ ……., but it can do several things better than a person can. The computer can't help you all 24/ ……., however. Several things need to be done before you will use it. 25/ ……. a computer programmer must tell the computer what to do. Then someone must fill in the facts the computer needs to 26/ ……. its work. Computers may seem to be smart, but they really aren't.

20. a/ since; b/ in fact; c/ nevertheless; d/ however
21. a/ use; b/ will; c/ do; d/ have
22. a/ have never taken; b/ have taken; c/ are taking; d/ will never take
23. a/ tool; b/machine; c/ device; d/ equipment
24. a/ by yourself; b/ on your own; c/ by itself; d/ on itself
25. a/ at the beginning; b/ first of all; c/ initially; d/ mainly
26. a/ make; b/ have; c/ do; d/ complete
27. a/ unless; b/ otherwise; c/ when; d/ if

Exercise 4. Read the text and decide which answer A, B, C or D best fits each space. Circle your answers.

Hideo Noguchi was born in 1876 in Japan. As an infant, he received a severe burn. 28/ ……., the Noguchi family was poor and couldn't afford a doctor for the child. 29/ ……., his left hand became paralyzed and deformed. Not until Noguchi was in his early teens 30/ ……. to a clinic where he 31/ ……. by Dr. Kane Watanabe. The operation and a series of treatments eventually 32/……. motion to the boy's thumb and little finger. To repay Dr. Watanabe, Noguchi worked summers at the clinic. During this time he saw the suffering of many patients and began to think about the 33/ ……. of helping them. 34/ ……. working at the clinic, Noguchi read all the doctor's medical books. Noguchi worked, studied and saved money to go the medical school in Tokyo. When he got there, he 35/ ……. a job as a janitor to support himself. Eventually, after much work he received his degree. In 1900 Dr. Noguchi came to America, where he 36/ ……. research on snake venoms. He later wrote an outstanding book on his topic. As a 37/ ……. member of the Rockefeller Institute for Medical Research, he spent many years studying the causes of diseases. In 1929 he went to Africa to study the causes of yellow fever, caught the disease and died.

28. a/ since; b/ however; c/ because; d/ on the other hand
29. a/ finally; b/ of which the result; c/ as a result; d/ eventually
30. a/ he had gone; b/ had he gone; c/ he went; d/ did he go
31. a/ he was operated; b/ he was operated on; c/ he had been operated; d/ he had been operated on
32. a/ brought back; b/ brought on; c/ recovered; d/ cured
33. a/ opportunity; b/ chance; c/ possibility; d/ occasion
34. a/ during; b/ while; c/ as; d/ since
35. a/ applied; b/ asked; c/ took; d/ took on
36. a/ wrote; b/ worked; c/ did; d/ made
37. a/ job; b/ stuff; c/ staff; d/ work

Exercise 5. Fill in the gaps in the following passage with the most suitable. Use only ONE word in each space.

In Poland, most jazz musicians start their careers 38/ ……. playing in students' clubs 39/ ……. the country. Sooner or 40/ ……. they enter a festival 41/ ……. the one at Wrocław. 42/ ……. they stand a chance of 43/ ……. both a large sum of money and a recording contract. The 44/ ……. step is to achieve international 45/ …….. Some artists become famous by concentrating 46/ ……. making records, 47/ ……. by touring abroad. In some cases they may even choose to live and work 48/……. permanently, as Michal Urbaniak has done in the United States.

Answer key

1/b; 2/a; 3/b; 4/c; 5/b; 6/a; 7/c; 8/c; 9/b; 10/b; 11/c; 12/ had come; 13/ took; 14 /could play; 15/ were shouting; 16/ cheering; 17/ wished; 18/ had; 19/ being; 20/ b; 21/ c; 22/ d; 23/ b; 24/ c; 25/ b; 26/ c; 27/ a; 28/ b; 29/ c; 30/ d; 31/ b; 32/ a; 33/ c; 34/ b; 35/ c; 36/ c; 37/ c; 38/ from/ with / by; 39/ around /in/throughout/ across; 40/ later; 41/ like; 42/ where /and; 43/ willing; 44/ next/ second/ third/ last; 45/ fame; 46/ on; 47/ others/ and/ or; 48/ abroad

REFERENCES

[1] Alderson, J.C., Clapham, C. & Wall, D., *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press, 1995.
[2] Aulia, I.F., Sukirlan, M.& Sudirman, "Analysis of the quality of teacher-made reading comprehension test items using Iteman," *U-JET* 3 (4), 2014.
[3] Bachman, L.E., *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press, 1990.
[4] Bachman, L.F. & Palmer, A.S., *Language Testing in Practice*. Oxford: Oxford University Press, 1996.
[5] *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Council of Europe, 2001.
[6] Crocker, L. & Algina, J., *Introduction to Classical and Modern Test Theory*. Chicago, Ill.: Holt Rinehart Winston, 1986.
[7] Defty, C. & Kusiak, M., "The practical English test," *Language Testing Update* 21, 1997, pp. 31-34.
[8] Hughes, A. *Testing for Language Teachers*. Cambridge: Cambridge University Press, 1989.
[9] Kusiak, M. & Jurek-Kwiatkowska, L.,"Towards standardisation of the final practical English examination in the Kraków cluster of colleges," *Innovations and Outcomes in English Language Teacher Education* , P. J. Melia, Ed. Warsaw: British Council, 1998, pp. 231-240.

AUTHORS

First Author – Kusiak-Pisowacka Monika, dr hab., The Institute of English Studies, Jagiellonian University, Kraków, Poland, monika.kusiak@uj.edu.pl.

Correspondence Author – Kusiak-Pisowacka, Monika, monika.kusiak@uj.edu.pl.