overSEAS 2016

This thesis was submitted by its author to the School of English and American Studies, Eötvös Loránd University, in partial fulfilment of the requirements for the degree of Bachelor of Arts. It was found to be among the best theses submitted in 2016, therefore it was decorated with the School's Outstanding Thesis Award. As such it is published in the form it was submitted in **overSEAS 2016** (http://seas3.elte.hu/overseas/2016.html)

ALAPSZAKOS SZAKDOLGOZAT

Káli Dominika

Anglisztika alapszak

Angol szakirány

2016

CERTIFICATE OF RESEARCH

By my signature below, I certify that my ELTE B.A. thesis, entitled **The predictive value of proficiency testing** is entirely the result of my own work, and that no material is included for which a degree has previously been conferred upon me. In my thesis I have faithfully and exactly cited all the sources (printed, electronic or oral) I have used, including books, journals, handouts and unpublished materials, as well as any other media, such as the internet, letters or significant personal communication, and have always indicated their origin.

Budapest, 15th April, 2016 Author

Káli Dominika

A HKR 346. § ad 76. § (4) c) pontja értelmében:

„… A szakdolgozathoz csatolni kell egy nyilatkozatot arról, hogy a munka a hallgató saját szellemi terméke…”

NYILATKOZAT

Alulírott **Káli Dominika** ezennel kijelentem és aláírásommal megerősítem, hogy az ELTE BTK **anglisztika** alapképzés/alapszak **angol** szakirányán írt jelen szakdolgozatom saját szellemi termékem, melyet korábban más szakon még nem nyújtottam be szakdolgozatként/záródolgozatként és amelybe mások munkáját (könyv, tanulmány, kézirat, internetes forrás, személyes közlés stb.) idézőjel és pontos hivatkozások nélkül nem építettem be.

Budapest, 2016. április 15.

.....

aláírás

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

Bölcsészettudományi Kar

ALAPSZAKOS SZAKDOLGOZAT

Az idegen nyelvi mérés előjelző értéke

The predicitve value of proficiency testing

Témavezető:

Dávid Gergely

adjunktus

Angol Nyelvpedagógiai Tanszék

Készítette:

Káli Dominika

Anglisztika alapszak

Angol szakirány

2016

Abstract

This study explores the predicitve value of the proficiency tests used at Eötvös Loránd University with respect to the BA thesis by comparing students' proficiency exam results to their theses grades. After providing an insight into the most significant pieces of literature related to the topic, including the notion of communicative language teaching and communicative competence, and the qualities of a good language test, with special emphasis on test validation process, the findings are presented using the pattern of mixed method research. The quantitative data analysis demonstrates the correlations between the grades, while the qualitative research involves some interviews with students, who share their experience about the proficiency exam and the thesis. Furthermore, these two documents are analysed, which helps to interpret the results and draw conclusions. The last part of the study explains the possible limitations of the research.

Table of contents

1. **Introduction** 1
2. **Literature review** 2
2.1. Communicative language teaching and testing 2
2.1.1. Communicative language teaching 2
2.1.2. Communicative language testing 4
2.2. The qualities of a good language test 5
2.2.1. Validity 5
2.2.2. Reliability 8
2.2.3. Test construction 10
2.2.4. Conclusion 11
3. **Research design** 11
3.1. Research questions 11
3.2. Research methods 12
3.3. The validation of the conclusions 14
4. **Results and discussion** 15
4.1. Quantitative analysis 15
4.1.1. Correlations 15
4.1.2. Interpretation of the correlations 15
4.2. Qualitative analysis 16
4.2.1. Results of the interviews 16
4.2.2. Discussion 18
4.2.3. Document analyses 18
4.2.4. Interpretation of the document analyses 20
5. **Conclusion** 20
5.1. Summary 20
5.2. Validation issues 21
5.3. Further possibilities of the research 22

References 23

Appendix A 25

Appendix B 27

1. Introduction

Language testing is an important part of language teaching, as it provides feedback for students about the level of their knowledge and helps them recognise what they should improve (Harmer, 2015). It is also useful for teachers, since tests reveal what are the weaknesses of the students and what they need to practise more. Each test is written for some purpose, and teachers must always be aware of what they are testing. For example, the objective of some tests is to foresee how students cope with a future challenge. Proficiency tests are often used for this purpose (Alderson, Clapham & Wall, 1995).

This paper aims to examine the predictive value of the BA proficiency exam of Eötvös Loránd University with respect to the BA thesis. My thesis follows up the suggestion that students who fail or perform poorly on the proficiency test of the university are more likely to have difficulties with the writing process of the thesis and receive bad grades than the ones who pass the proficiency exam with a good result (Dávid, Gergely, personal communication, 8 December, 2015). The main question of the research is whether the results of these assessments correspond to each other. It will also be useful to examine if the parts of the proficiency exam, the Use of English paper and the oral exam, correlate differently with the theses results. Finally, this paper looks for the reasons that explain the final outcome of the study.

I believe that the question of my research is both an interesting and a debatable one, since in spite of the fact that the proficiency exam and the thesis do not seem to measure the same language skills, it may be possible to identify abilities measured by both of them. Relying on the test specifications, the first part of the BA proficiency exam "focuses mainly on grammatical, lexical and discoursal accuracy" (Information on the BBN ANG-001/ANG002, n.d., para.1), while the oral part of the exam requires fluency and accuracy,

along with the appropriate use of grammar and good communication skills. On the other hand, being able to write a thesis involves completely different abilities, such as satisfactory writing and synthesising skills or "using appropriate methodological and stylistic apparatus" (Quality of Thesis, n.d., para. 1). It would therefore be a remarkable observation if the proficiency grades were proved to be reliable predictors of the theses grades.

After providing a review of the relevant literature to present the educational background of the topic, I want to gather information about students' results of their proficiency exams and theses, and make calculations to discover how the grades correlate to each other. On the basis of these correlations, I will also make interviews with some of the students to be able to access their personal background and make my research more extensive, analysing their own experience. Furthermore, both assessments will be analysed to explore what kinds of abilities they measure.

I expect that this study will prove to be a useful investigation, and its results will define how close the connection is between the two kinds of assessments. If the correlations are high, they may trigger further research in this area. In the case of a contradictory outcome, the study affirms that predictive validity cannot be established and students' performance on the proficiency exam does not truly predict their theses grades.

2. Literature review

2.1. Communicative language teaching and testing

2.1.1. Communicative language teaching

In the 1960s, students were considered to be successful language learners if they could create grammatically correct sentences in the target language; however, the end of the

following decade brought significant changes in the field of language teaching, and Hymes was that prominent figure whose theory deeply influenced the linguists of this new period (Johnson, 1981). Although he partly agreed with the structuralism of the 1960s, he thought that the knowledge of grammar was not sufficient to communicate effectively in a language. His idea was that successful communication also "involved the ability to be appropriate, to know the right things to say at the right time" (Johnson, 1981, p.2). This ability was called communicative competence. Nowadays, it is almost impossible not to hear about the widespread concepts of communicative teaching and communicative competence (Harmer, 2015).

It is difficult to define what communicative language teaching (CLT) exactly means, as even in a standard methodological book, Harmer (2015) acknowledges the many possible interpretations of CLT. However, most people agree that this new teaching method focuses on the content of what students produce in a language class, and the specific language forms are of minor significance (Harmer, 2015). As Littlewood (1981) states, CLT "pays systematic attention to functional as well as structural aspects of language" (p.1), so not only communication, but language as a means of communication has an important role in communicative teaching, too.

CLT and communicative competence are strongly related to each other, as the aim of communicative language teaching is to develop students' communicative competence (Johnson, 1981). The most acknowledged model of communicative competence has been framed by Canale and Swain. The model consists of four components: the linguistic, the sociolinguistic, the discoursal and the strategic competences (Weir, 1990). The linguistic competence means the appropriate use of grammatical rules, the sociolinguistic competence assumes the knowledge of sociocultural rules, the discoursal competence is concerned with the cohesion and coherence of an utterance and the strategic competence means the ability to

use different communication strategies (Weir, 1990). Improving each of the components of the communicative competence became the main target of communicative language teaching.

2.1.2. Communicative language testing

The new communicative approach has changed not only language teaching but also language testing. When teachers started to follow the principles of communicative language teaching, they realized that this new method should involve a new kind of testing, as well. According to Morrow (1979), there were several issues that needed to be revised in earlier tests in order to be able to assess students' communicative proficiency. One of them concerned Lado's structuralist approach, which says that discrete elements of language should be taught and tested; however, learners should also know how to put them together in different situations because it is a part of communicative competence. Another problem comes from the behaviourist evaluation of answers, since the drilling of students and habit formation are not fundamental parts of communicative testing. In communicative language testing, answers "are more than simply right or wrong" (Morrow, 1979, p.145). The last point raised by Morrow (1979) is the complex relationship between reliability and validity, two concepts that have a great role in writing a good communicative language test.

Apart from the examples discussed above, test writers face many different kinds of difficulties during the writing process; however, it is worth investigating what makes a good communicative test, since this paper focuses on the BA proficiency exam, which does not seem to assess students' communicative competence, as it covers a lot less ground than it should. This point will be discussed at the validation issues of the research. Besides assessing students' communicative competence, there are three important components, which should be considered as the basics of a good communicative language test, and they should never be

ignored when writing a test: validity, reliability and the well-planned process of test construction.

2.2. **The qualities of a good language test**

2.2.1. Validity

Validity provides answer for an important question: "Does the test test what it is supposed to test?" (Alderson, Clapham & Wall, 1995, p.170). For instance, if teachers want to measure students' knowledge of English grammar, a test containing questions about Great Britain's history would not be valid for that purpose. This history test can provide an accurate measure of a student's historical knowledge, but its result would not help English language teachers to decide if their students are familiar with the English grammar. Consequently, it is important to emphasize that it is not the test that must be valid but the inference based on the results of the test (Messick, 1995).

The interpretation of validity has changed since the 1980s. Following the traditional perception, Hughes (1989) defined five different kinds of validities: content validity, criterion-related validity, construct validity and face validity. *Content validity* is one of the most important aspects of a test, since "the greater the test's content validity, the more likely it is to be an accurate measure of what it is supposed to measure" (Hughes, 1989, p.22). If it does not develop the skills written in the test specifications, it cannot have content validity. *Criterion-related validity* means comparing the results of the tests with a previously determined criterion measure which is believed to be a reliable measurement of the ability tested (Bachman, 1990). For example, at an oral examination, the performance of a candidate is graded with the help of a standardized document, containing all the structures and functions that they need to be able to use properly. A subtype of criterion-related validity is called

predictive validity, which examines if a test score can predict students' future performance. Predictive validity belongs to the group of the empirical kinds of validity, which means that it "involves the collection of data or recourse to mathematical formulae for the computation of validity coefficients" (Henning, 1987, p.94). Examining predictive validity will have major significance in the research design of this thesis. Bachman (1990) highlights that it is problematic to rely on predictive validity, as it often "ignores the question of what abilities are being measured" (p.250). It is therefore necessary to know if the test deals with the right abilities. This question may be answered through the construct validation of the test. *Construct validity* ensures test users that the test measures "just the ability which it is supposed to measure" (Hughes, 1989, p.26), so no other ability is tested in a test or in a part of a test. "The purpose of construct validation is to provide evidence that underlying theoretical constructs being measured in a language are themselves valid" (Henning, 1987, p.98). Finally, *face validity* is about the outward aspect of a test. It is not considered to be meaningful in the validation process, because it is often based on the opinion of non-experts. However, some think that face validity is worth dealing with, as students are more likely to have better results if they are convinced that the test has validity (Alderson, Clapham & Wall, 1995).

While Hughes (1989) has made a distinction between different types of validities, Alderson, Clapham and Wall (1995), arguing against the traditional view, suggest that validity is a complex term and these types listed by Hughes should rather be called as different ways of determining validity. Messick (1995) and Bachman (1990) also agree that validity is a unitary concept, and those categories used in the traditional interpretation are not the types of the validity but the "complementary types of evidence that must be gathered" (Bachman, 1990, p.243) in order to make the inference of a test valid. They put the emphasis on the evidential basis of validity, which means that in their point of view, the main concern of establishing validity is gathering satisfying evidence. Messick (as cited in Brown & Hudson,

2002) compares the validation process to a trial, where the lawyer needs to defend the client, supporting his case with convincing evidence. He defines validity as "an argument, or more often a series or arguments, for the effectiveness of a test for a particular purpose" (Brown & Hudson, 2002, p.240-241).

In the light of what Messick (1995) and Bachman (1990) claims, evidence plays the most important role in test validation. The first step of the validation process is determining what purpose the test is valid for, and the second one is to collect enough evidence to justify that there is connection between the results and the use of the test (Bachman, 1990). A validity argument can hold and deal with a number of different kinds of evidence. The three main groups of evidence defined by Bachman (1990) are: content relevance (content validity), criterion relatedness (criterion-related validity) and meaningfulness of construct (construct validity). Alderson, Clapham, Wall (1995) and Bachman (1990) agree that the different kinds of evidence can strengthen or weaken validity.

The best way to strengthen it is to collect as many types of evidence as possible, as "the more different 'types' of validity that can be established, the better, and the more evidence that can be gathered for any one type of validity, the better" (Alderson, Clapham & Wall, 1995, p.171). Consequently, it is essential to examine a test from several aspects, since each kind of evidence lacks something important. For example, content relevance is one of the most important kinds of evidence in the validation process; however, other 'types' of evidence should be gathered too, as content relevance itself is not convincing enough to call the inference of a test valid. It does not support any conclusion about language abilities and does not examine the performance of test takers (Bachman, 1990). Furthermore, the evidence of criterion relatedness proves that test scores validly indicate a specific language ability (Bachman, 1990), but it ignores the fact that they could be indicators of other abilities, too.

According to Messick (1995), only construct validation can serve as a good evidential basis of validity, as it involves content- and criterion-related evidence, as well. Bachman (1990) defines construct as the "definition of abilities that permit us to state specific hypotheses about how these abilities are or are not related to other abilities" (p.255). However, problems might arise when establishing construct validity. For instance, 'construct underrepresentation' occurs when "the assessment is too narrow and fails to include important dimensions of the construct" (Messick, 1995, p.742). The opposite of construct underrepresentation is the construct-irrelevant variance, which means that the assessment is too extensive and it contains several components which make a task irrelevantly difficult or easy (Messick, 1995). An example of Messick (1995) perfectly demonstrates both cases. If a test intends to measure one's reading comprehension skills, it should not consist of questions which are not answered in the text but they assess background knowledge, since that would make the task irrelevantly difficult. On the other hand, if the content of a test is well-known by the student, it is easy to understand it and answer the questions without reading the text carefully. Construct underrepresentation and construct-irrelevant variance usually lead to invalidity (Messick, 1995). In conclusion, each type of evidence helps establishing validity but separately they do not guarantee as strong validation arguments as they could do so together. Pitfalls may occur during the process of construct validation, as well.

2.2.2. Reliability

Another important quality of a good language test is reliability. A test is reliable if it measures students' ability consistently. To put it as simply as possible, there would not be a huge difference between the results of a test if it was administered a large number of times (Hughes, 1989). The reliability of a test can be measured with the comparison of the test scores. It is called reliability coefficient, which should be 1, if two sets of scores are equal.

Alderson, Clapham, Wall (1995) and Bachman (1990) claim that there are a lot of influential factors which might have an adverse effect on the performance of the students; it is therefore not possible to write a perfectly reliable test, since it would be hardly achievable for students to perform equally on two different occasions. Hughes (1989) adds that reliability coefficient also depends on which ability of a student is measured and how serious decision is made on the results of the test. Although several factors can affect students' scores and the height of this coefficient, test writers should strive to design highly reliable tests because if test scores are absolutely different from each other, teachers would never know which score reflects the real performance (true score) of a student (Hughes, 1989). What is more, reliability is an extremely influential factor in establishing the validity of a test.

Validity could not be undermined by simply a particular measurement not being valid but it is undermined e.g. through reliability problems. Validity and reliability are connected to each other, and this connection is so close that it is incredibly difficult to separate the two notions (Hughes, 2003). There is a kind of hierarchy between these concepts. The conclusions based on a test "cannot be valid unless the test is reliable" (Alderson, Clapham & Wall, 1995, p. 187). Although it is attractive to say that if a test is reliable, it is valid at the same time, a test can be consistently wrong. On the other hand, it is also popular to think that a test may not be very reliable but it is certainly highly valid. Nevertheless, it is important to see that reliability or validity itself is not enough to establish in a good test. Reliability is a precondition for validity, but validity is as essential as reliability. The main question for language testers: which is the more important? Teachers' thinking ranges between these two extremes; however, most interpretations of the tests are valid for some purpose and it depends on the context how high the reliability and the validity coefficient should be. Hughes (1989) states that "the tester has to balance gains in one against losses in the other" (p.42), depending on the purpose of the test. Heaton (1988) highlights that if a test contains a lot of

characteristics which make it reliable, these features will reduce its validity, as well. It is therefore "essential to devise a valid test first of all and then to establish ways of increasing its reliability" (Heaton, 1988, p.165).

2.2.3. Test construction

Last but not least, the careful construction of a test is also an influential factor that contributes to the success of a test. Many think that constructing a language test is a quick process; however, it requires a lot of effort, and a good test goes through several steps before it is used in the classroom. Alderson, Clapham and Wall (1995) have identified the different stages of test construction. Firstly, a good test has specifications, which help test writers and users to choose which test is the best for their own purposes. These pieces of information may be carefully elaborated or contain only the most relevant points of the test, depending on the need it should fulfil. It also serves as guidance for determining the content validity of the test. The next step is item writing and moderation, which should be done by well-qualified experts, who are aware of the expectations and have some teaching experience and ingenuity. When the items are ready, they need to be moderated by another expert who can form opinion about them and change them if it is necessary. Alderson, Clapham, Wall (1995) and Hughes (1989) agree that moderation is an extremely important part of item writing, since it is useful to examine items from different points of view and develop them. After moderating test items, a draft test paper is constructed, which goes to a formal committee of experts, who take the test as they were students. Alderson, Clapham and Wall (1995) refer to this stage as the "main trial" before the tests are written in a real classroom situation, while Hughes (1989) says that before the test are used, pretesting can be done by a group of students which is similar to the target group. Finally, results are analysed and further changes may be made to improve the tests.

2.2.4. Conclusion

All in all, in order to write a good communicative language test, test writers should establish different ‘types’ of validity supported with sufficient evidence, make the test as reliable as possible, and pay attention to the process of test construction. Validity seems to be the most important quality of a test, since tests must always demonstrate that they measure those abilities which are intended to measure by the teachers. It is extremely useful to present both the pro and the contra arguments of determining a particular type of validity, since it depicts a full picture of the difficulties of validation, and makes it clear that validating any kind of test is a complex process. The inference from a test can only be valid, if the validity arguments are supported by adequate evidence. Collecting only one kind of evidence may weaken a test’s validity. Among the different ‘types’ of validity, construct validation seems to be the best evidential basis of validity, but construct underrepresentation or construct-irrelevant variance can still engender invalidity. Besides the lack of evidence and the one-sided validation of a test, reliability issues should not be neglected either, since validity would not exist without reliability. Finally, the careful construction of a test also contributes to the success of the test, as the more people work on a test, the more efficient and reliable it will be.

3. Research design

3.1. Research questions

Focusing on the validation of an interpretation on the basis of predictive evidence, this research aims to establish the extent to which students’ results of the proficiency exam can predict the grades of their theses. Answering this question demands further observations on the topic. Firstly, as there are no similar studies available in this particular issue, it is

extremely difficult to forecast if the outcome of the research will result in high or low correlation between the grades. Both options need to be taken into account; therefore, I approach this issue with an open mind, and I do not form an initial hypothesis about the results. Secondly, the proficiency exam consists of two parts, and it might be useful to examine those parts separately and compare them to the theses grades, since the scores of the Use of English test and the oral examination may correlate differently with them. Finally, it is also necessary to discover the reasons why students' performance on the proficiency exam differ from or coincide with their theses grades.

3.2.Research methods

To provide answers for all the questions raised above, the pattern of mixed methods research will be used (Creswell, 2014), containing both quantitative and qualitative analysis of the data. The reason why this kind of method has been chosen is that it provides a wider picture about the issue and represents more than one point of view. Creswell (2014) says that one of the greatest advantages of mixed methods research is that it leads to a deeper analysis of the matter and makes the research more interesting.

As for the results, they will be introduced in the style of 'explanatory sequential mixed methods design'. It means that the study begins with the analysis of the quantitative data, and "then builds on the results to explain them with qualitative research" (Creswell, 2014, p.15). It is important to emphasize that in the explanatory sequential mixed methods design, the qualitative part of the research supports the quantitative data, which makes the research more reliable and valuable.

For the quantitative part of the study, the grades of the proficiency exam written in 2012 are collected and compared to the grades of the BA theses of the same students. The theses were submitted in the spring or autumn term in 2014. Correlation coefficients will be

calculated not only between those marks, but also between the oral examination and the thesis, and the Use of English test and the thesis.

The qualitative research will include the interpretation of ten interviews with individuals to present a better understanding of the topic. Semi-structured interviews were selected as the method of the qualitative data collection, as they contain set questions, which help researchers to analyse the answers, but the students can also share any relevant information which might not be included in the set questions. In this way, semi-structured interviews bring respondents' thoughts, feelings and attitudes closer to the researcher, allowing a deeper analysis of the topic (Barriball & While, 1994).

Not being able to draw a large sample for the interviews, I resorted to establishing patterns of results, in order to present the greatest possible range of opinion. Five categories were made, and each category consisted of two students. In the first category, students received excellent marks for both the proficiency exam and the thesis. In the second one, their results were the same, either a 4 or a 3, for both tests. The third category involved students whose proficiency grades were excellent, but the grades of their theses were weaker, either 4 or 3. The fourth pair of students was the opposite of the previous category, as at first they failed the proficiency exam, but their theses grades were excellent. The last group contained students who performed poorly (received 1 or 2) in both cases.

The same questions were asked of all the students, and the questions required mainly long reasoning from the candidates, expressing their own opinion in connection with the topic of the research. The interviews were conducted in Hungarian, since speaking in English may limit the accuracy and the amount of information shared by the respondents. Due to the potentially sensitive nature of the research, the identities and other personal data of the participants have been protected. (For the sample of the Hungarian interview questions, see Appendix B).

Finally, using the test specifications and the criteria found on the website of the Department of English Language Pedagogy and the School of English and American Studies at Eötvös Loránd University, document analyses were made to investigate what kinds of skills are required to achieve a good mark in the Use of English test, the oral exam, and the thesis.

3.3.The validation of the conclusions

In order to draw valid conclusions on the basis of the correlations, it is essential to collect satisfactory evidence to support the numerical data with valid arguments. As predictive validity itself is not sufficient to validate the interpretations, establishing construct validity will also be important, since "one cannot claim that a test has criterion-related validity because it correlates highly with another test, if the other test itself does not measure the criterion in question" (Weir, 1990, p. 28). If the correlations are high, there must be similar constructs in both assessments; however, in the case of low correlations, they will prove to be tests of different abilities. As a consequence, both the proficiency test and the thesis need to be analysed to discover which abilities are measured by each one, and compare them to each other. Furthermore, the subjects of the interviews will also contribute to the validation process, since their observations will be used to establish face validity. The in-depth document analyses and the qualitative part of the research together will strengthen the validity of the conclusions from the whole investigation.

4. Results and discussion

4.1. **Quantitative research**

4.1.1. Correlations

In the quantitative data analysis of this study, four correlations were made (see Chart 1 and 2 in Appendix A). The first correlation coefficient was counted between the grades of the proficiency exams and the theses. Surprisingly, the correlation proved to be drastically low, not more than 0,28 ($r_{Pearson}(110) = 0,28$, $p< 0,01$). It is worth mentioning that only from the proficiency grades it is not possible to know whether students' scores were closer to the lower or the upper grade boundaries. A wider picture is provided through the examination of the total score of the proficiency exam, since these scores can show the differences between the same marks. This correlation, calculated between the total scores and the theses grades, is 0,42 ($r_{Pearson}(101) = 0,42$, $p< 0,01$), which is the highest coefficient among the results. The last two correlations investigated the connection between the theses grades and the parts of the proficiency exam separately. The Use of English test and the thesis show a correlation coefficient of 0,29 ($r_{Pearson}(110) = 0,29$, $p< 0,01$), while the coefficient of the oral exam and the thesis is 0,24 ($r_{Pearson}(110) = 0,24$, $p< 0,01$), which is the lowest one of all.

4.1.2. Interpretation of the correlations

The investigation of the correlation coefficients clearly suggests that the students' proficiency grades are merely different from the results of their thesis. There is only a little correspondence between these assessments. Although the correlation calculated between students' total scores and their thesis grades is moderately strong with respect to the other ones, neither of the correlations is too high, which leads to the conclusion that the proficiency exams cannot truly predict one's thesis grade. As the quantitative research does not provide an

explanation for the outcome of the calculations, it is useful to demonstrate the results of the interviews, which reveal what might be the reason of the low correlation coefficients.

4.2. Qualitative research

4.2.1. Results of the interviews

Firstly, students were asked about what kinds of abilities are necessary to receive good marks for the proficiency exam and the thesis. According to the respondents, the oral exam requires satisfactory communication skills and a decent lexical knowledge, which enables students to express themselves in many different ways. Fluency is also thought to be an important criterion of the exam, as minor grammar mistakes will not decrease one's points at the oral exam so much as the lack of fluency would do so. The use of English test is associated with the same abilities in each student's mind: a stable knowledge of English grammar and being familiar with common English phrases and idioms. On the basis of their answers, the wide range of vocabulary is also essential in this part of the exam. In contrast to the proficiency exam, the focus of the thesis is considered to be on other language skills. Most of the respondents emphasized that thesis writing included almost all of the skills required for the proficiency exam; however, they are of minor importance. The most significant abilities they listed are the use of the academic register, the knowledge of that particular jargon that one's topic demands, and being able to create a coherent, longer piece of research, which is well-constructed and easy to follow. Some students also added that writing a thesis does not concentrate on the use of English language, since that should not mean any problem for students by the end of the third year. However, respondents agree that grammatical knowledge and the range of vocabulary partly contribute to the success of the thesis.

Moreover, students were asked if they could formulate a hypothesis about whether the correlations are high or low between the grades, and explain their choice. 7 out of 10 students

do not expect high correlation, as they think that the proficiency exam and the thesis are not similar to each other. The explanation for their answer is that not the same languages skills are measured in each one, and the forms of the tests are different as well, since the Use of English test and the oral exam require instantaneous reaction from the students, while in the case of thesis writing, students can work on their research during several months, and edit it as many times as they want to. On the other hand, three interviewees believe that the correlations may be high, and they highlight that the oral part of the proficiency exam might be the best predictor of one's thesis grade, as the written and the spoken language often coincide in some respects. They therefore think that the knowledge of vocabulary and grammar are crucial requirements in both.

Students also listed their strongest and weakest language skills, and their own difficulties in each exam. These two questions were supposed to be closely related to each other; however, results clearly show that the respondents think that their grades were largely influenced by other factors (e.g. language anxiety, the persona of the supervisor and the opponent, or the choice of the topic) rather than their language skills. These answers therefore do not support the construct validation of the research, as pupils strongly believe that their given marks were affected by various external factors, but not by the lack of certain language skills.

The last question examined whether students can explain why their grades of the proficiency exam and the thesis were noticeably similar or different. The majority of the respondents firmly agree that the results are independent from each other. They claim that the proficiency exam may predict the scores of some particular parts of the thesis, but its influence on it is insignificant. Instead, most of the interviewees consider the academic writing class offered at the university a truly reliable predictor of the thesis grade, since the

skills acquired during those lessons (e.g. synthesising, summarising) are the ones which are needed for thesis writing.

4.2.2. Discussion

The results of the qualitative research seem to strengthen the validity of the conclusions drawn from the correlations. Despite selecting extremely different patterns of results, expecting them to reveal considerable differences between students' thoughts and attitudes, the answers show that the majority of the students expect the same results from the research.

In order to explore whether the observations of the interviewed students provide a real picture about the proficiency exam and the thesis, a detailed document analysis is required about both of them, which reveals which constructs are measured in each one. If there is a huge contrast between the abilities determined in the tests, the analyses will strengthen the validity of the research, making the conclusions of the interviews more valid.

4.2.3. Document analyses

The first part of the proficiency exam is a Use of English paper, which is made up of 75 questions, and it is 90 minutes long. Each multiple-choice item is 1 point, while answers requiring filling in a longer string of words are 2 points. Students do not need to produce long, written texts, but reading comprehension skills are necessary, as the context of a short passage can often help to fill in the gaps or answer a question. "The requirements of the exam are based on the Common European Framework (CEF)" (Overview of the BA Language Proficiency Exam structure, n.d., para. 2); therefore, it is a criterion-referenced test. Kontra and Kormos (2007) state that CEF helps "describing foreign language competence in a uniform way" (p.17) and also facilitates drawing up test specifications. According to the test

criterion, what are needed for the proficiency exam are the knowledge of a "wide range of complex grammatical structures and advanced vocabulary, and the advanced use of discourse" (Specifications for BA Language Proficiency Exam, n.d., para. 1).

The Oral part of the exam is done in groups of three, and it takes around 25 minutes. It consists of three parts. In the first five minutes, students are allowed to prepare for the topic they need to discuss. This section is not marked by the examiners. The first phase which is marked necessitates candidates to speak individually and help each other with some questions, while in the second phase, students are involved in a conversation where they should actively communicate with each other. The oral examination is also a criterion-referenced test, based on the CEF, and the test specifications are the same as the ones defined in the Use of English paper; however, the evaluation of the answers is carried out on the basis of the following criteria: "fluency and production, content, range and flexibility, accuracy and interaction" (BA Oral Proficiency Criteria, n.d.).

The results of the analysis of the thesis reflect how different it is from the proficiency exam. The thesis is "a longer piece of scientific research" (Quality of thesis, n.d., para.1), which requires students to write at least 40000 characters (including spaces) about a topic of their choice, and present their first academic research paper, using "the appropriate methodological and stylistic apparatus required in English academic settings" (Quality of thesis, n.d., para.1). It is also important to read books and articles in the field of the students' choice, and incorporate this knowledge into the thesis. Writing and synthesising skills, along with the use of academic style, are the main concerns of thesis writing. Its grade consists of the evaluation of academic achievement and language competence, so grammar, spelling and punctuation are also parts of the assessment.

4.2.4. Interpretation of the analyses

After the detailed investigation of each document, it is evident that the proficiency exam and the thesis measure different kinds of constructs; however, language competence is assessed by both of them. Comparing the results of the interviews to the content of the document analyses, the respondents seem to be aware of what kinds of skills are measured in each assessment, since the majority of their answers almost exquisitely describe the tests' criteria.

5. Conclusion

5.1. Summary

Taking each part of the research into consideration, it is clear now that the BA proficiency exam is not a reliable predictor of the BA thesis grade. Besides the low correlation coefficients, the answers of the interviewees and the document analyses also justify that these two kinds of tests are so different from each other that it is not possible to establish predictive validity. Without knowing how low the correlations were, students' responses perfectly support the results of the quantitative analysis, and what they said accurately summarize the content of the analyses of the tests. Although the advanced knowledge of English grammar and vocabulary is covered by both, these skills do not play as important roles in thesis writing as they do in the proficiency test. The interviewed students emphasized that they felt external factors more influential than the lack of some of their language skills.

I believe that the most important conclusion drawn from the results of this thesis is that although the proficiency grades do not highly correlate with the theses marks, they are able to show students' weaker and stronger language skills, and they may serve as a guideline for the supervisors, if they want to see what kind of difficulties each student might encounter during the thesis writing process.

5.2. Validation issues

Unfortunately, there are some issues which weaken the validity of the conclusions drawn from the research. For instance, in the quantitative analysis, there are more proficiency grades than theses grades in this particular time interval. This phenomenon is called the 'truncated sample problem', and it presumably lessens the predictive validity coefficient (Alderson, Clapham & Wall, 1995). The reason of the difference might be that a lot of students could not finish their subjects in time and decided to submit their theses in the subsequent term or they might dropped out from the course. In 2011, 306 students took the BA proficiency exam at the university; however, only 111 of them were able to submit their theses at the end of the third year. It means that approximately one third of the data available has been used in the analysis. Besides the missing theses grades, nine students have not achieved a total score in the proficiency exam, as those who failed either the Use of English test or the oral part of the exam could not pass the proficiency exam, and their total scores were not counted. Despite dealing with a truncated sample, the data available for the research was still sufficient to draw general conclusions and to provide an insight into the most significant tendencies.

Moreover, the number of the selected students for the interviews may seem insufficient, as it is too small to derive far reaching conclusions; however, the interviews seem credible to me, since towards the end of the interview sequence, the respondents did not share substantial information with me, which reminds me of a possible saturation of data (Creswell, 2014). Probably, it would have been useless to conduct more interviews, since a general picture was provided by the small number of respondents, as well.

Finally, as the BA proficiency exam does not measure neither writing nor listening skills, it may not depict a reliable picture of students' communicative proficiency. Maybe the writing skills of some students are much better than their speaking skills, but since it is not

measured in the proficiency exam, students do not have the opportunity to give an account of this part of their knowledge. According to Hughes (1989), tests should be "long enough to achieve a satisfactory reliability" (p.37). If the test is not reliable, it will not be valid either. Adding more components to the exam might provide a more real image of pupils' communicative proficiency.

5.3. Further possibilities of the research

As the academic writing exam measures students' academic writing skills, it may have been practical to collect the participants' grades of the academic writing exam and make correlations with those marks, as well, to examine whether they are able to predict students' theses grades better than the proficiency tests, which would mean that the skills measured in the academic writing exam are more influential in thesis writing than the ones assessed in the BA proficiency exam.

References

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge, England: Cambridge University Press.

BA Oral Proficiency Criteria. (n.d.). Retrieved from http://delp.elte.hu/BAlangprofinfo.htm#Specifications_for_the_Oral_Test

Bachman, L. (1990). *Fundamental considerations in language testing*. Oxford, England: Oxford University Press.

Barriball, K. L., & While, A. (1994). Collecting data using a semi-structured interview: A discussion paper. *Journal of Advanced Nursing, 19*(2), 328-335.

Brown, J. D., & Hudson, T. (2002). *Criterion-referenced language testing*. Cambridge, England: Cambridge University Press.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches*. Los Angeles: SAGE Publications.

Harmer, J. (2015). *The practise of English Language Teaching* (5th ed.). Harlow, England: Pearson Education.

Heaton, J. B. (1988). *Writing English language tests*. London, England: Longman.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research*. Cambridge, MA: Newberry House.

Hughes, A. (1989). *Testing for language teachers*. Cambridge, England: Cambridge University Press.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge, England: Cambridge University Press.

Information on the BBN ANG-001/ANG002 Language Proficiency Exam. (n.d.). Retrieved from http://delp.elte.hu/BAlangprofinfo.htm

Johnson, K. (1981). Some background, some key terms and some definitions. In Johnson, K., & Morrow, K. (Eds.), *Communication in the classroom: Applications and methods for a communicative approach* (pp.1-12). London, England: Longman.

Kontra, E., & Kormos, J. (2007). *An introduction to language testing for teachers of English*. Budapest, Hungary: Okker Kiadó.

Littlewood, W. (1981). *Communicative language teaching: An introduction*. Cambridge, England: Cambridge University Press.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist*, *50*(9), 741-749.

Morrow, K. (1979). Communicative language testing: revolution or evolution? In Brumfit, C. J. & Johnson, K. (Eds.), *The communicative approach to language teaching* (pp.143-157). Oxford, England: Oxford University Press.

Overview of the BA Language Proficiency Exam structure. (n.d.). Retrieved from http://delp.elte.hu/BAlangprofinfo.htm#Overview_of_the_BA_Language_Proficiency_

Quality of Thesis. (n.d.). Retrieved from http://seaswiki.elte.hu/studies/BA/major/graduation/thesis/requirements

Specifications for BA Language Proficiency Exam. (n.d.). Retrieved from http://delp.elte.hu/BAlangprofinfo.htm#Specifications_for_BA_Language_Proficien

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

APPENDIX A

Chart 1

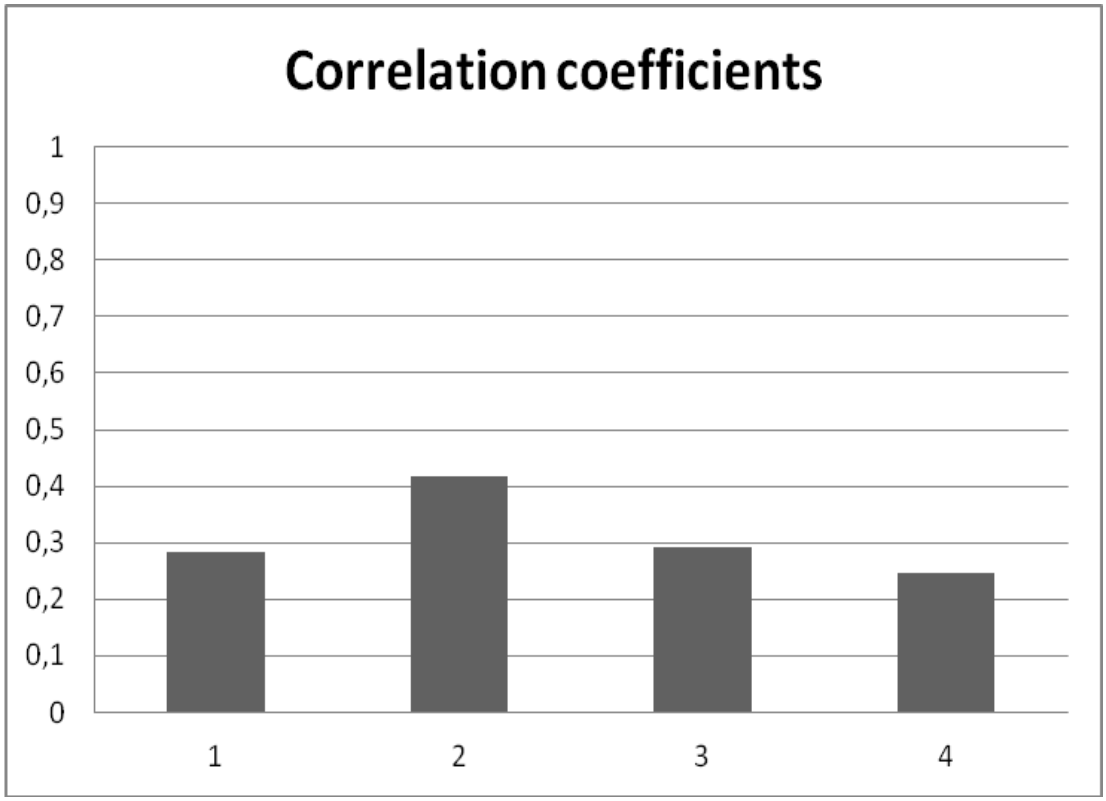
Correlations

		ORAL	UoE	GRAD	TOT	Thesis
ORAL	Pearson Correlation	1	,437**	,790**	,849**	,245**
	Sig. (2-tailed)		,000	,000	,000	,010
	N	111	111	111	102	111
UoE	Pearson Correlation	,437**	1	,806**	,800**	,291**
	Sig. (2-tailed)	,000		,000	,000	,002
	N	111	111	111	102	111
GRAD	Pearson Correlation	,790**	,806**	1	,960**	,284**
	Sig. (2-tailed)	,000	,000		,000	,003
	N	111	111	111	102	111
TOT	Pearson Correlation	,849**	,800**	,960**	1	,417**
	Sig. (2-tailed)	,000	,000	,000		,000
	N	102	102	102	102	102
Thesis	Pearson Correlation	,245**	,291**	,284**	,417**	1
	Sig. (2-tailed)	,010	,002	,003	,000	
	N	111	111	111	102	111

**. Correlation is significant at the 0.01 level (2-tailed).

(courtesy of Dávid Gergely, senior lecturer of ELTE)

Chart 2



1- Thesis grade - Proficiency grade

2- Thesis grade - Total score

3- Thesis grade - Use-of-English score

4- Thesis grade - Oral score

APPENDIX B

Név:

Nyelvi alapvizsga érdemjegye:

Szakdolgozat érdemjegye:

1. Mit gondol mely nyelvi készségek a legfontosabbak az alapvizsga teljesítéséhez?

2. Milyen nyelvi készségeket igényel a szakdolgozat megírása?

3. Lehet hasonlóság a kettő között? Kérem, indokolja meg válaszát!

4. Mely nyelvi készségeit érzi a legerősebbnek és a leggyengébbnek?

5. Ön szerint mik voltak az alapvizsga és a szakdolgozat nehézségei?

6. Mi lehet az oka annak, hogy a két eredménye nagyon megegyező/eltérő?