University of Szeged
Graduate School in Linguistics
English Applied Linguistics PhD Program

Establishing the Context and Scoring Validity of the Writing Tasks of Euroexam International's English for Academic Purposes Test

PhD Dissertation
Summary

Fűköh Borbála

Supervisor: Dr Barát Erzsébet

Szeged
2020

1. Introduction

High-stakes language testing plays an important role in the education system of Hungary. The language exams are supervised by the Accreditation Board for Foreign Language Examinations, which is responsible for standardising the professional requirements for examination boards across the country. Currently both major international test providers and locally developed exams are available for test takers (Educational Authority, 2020).

The major test providers, such as Pearson, IELTS and TOEFL, offer academic tests for students who desire to continue their studies in English language higher education (IELTS, 2018; Pearson PTE Academic, 2017; TOEFL iBT, 2018). Despite the growing number of Hungarian students pursuing university studies in European Union and UK universities, there are no state accredited English for Academic Purposes (EAP) exams available in Hungary yet. In 2017, Euroexam International decided to launch an exam development project to make up for this gap and designed an English for Academic Purposes (EAP) test targeted at Hungarian and East-Central European students (Euroexam Academic, 2019a). As member of the Euroexam Research Team as well as an ESP instructor in tertiary education, I undertook the task of leading the validation research for the writing tasks of the test (Fűköh, 2019a, 2019b). The stages of validation are to prove that the writing tasks of the Academic Exam of Euroexam International reflect the skills needed in the different academic discourse types students need to perform in the course of their learning.

The dissertation is divided into two main parts. The first part of the dissertation is a review of the relevant literature in four chapters. After introducing the background to the research, Chapter 2 brings together the relevant theoretical works on the nature of writing and the writing process with a special focus on the nature of academic writing. Chapter 3 is devoted to the topic of assessing writing. In addition to discussing the nature of writing assessment in general, and second language writing as it appears in the Common European Framework of Reference (Council of Europe, 2001; 2018), I focus on raters, rater leniency and harshness, and rater training. The last part of the chapter explores the use of rating scales and checklists, their advantages and disadvantages. In Chapter 4, I give an overview of test development, more specifically test validity. I discuss the development of language tests, more particularly the characteristics of test usefulness as presented by Bachman and Palmer (1996; 2010). The second part of this chapter is devoted to the various concepts of validity in language testing. I present the traditional and the new concepts of validity, and

discuss the socio-cognitive framework by Weir (2005). In addition, I discuss the concept of localisation (O'Sullivan & Dunlea, 2015) that is of particular relevance in the context of international university admissions.

Chapter 5 is focused on the methods of generating validity evidence and argue for the advantage of mixed-methods research in the test development process. This is the chapter that formulates and argues for the research questions of the dissertation in the context of my main area of research. I present my empirical research in the second part of the dissertation in three chapters. These chapters follow the stages in the test development process of the Euroexam Academic project. Chapter 6 outlines the initial development phase, in other words, the planning and domain analysis phases of the test development process as well as the reflection on the judgement of an external expert panel. Chapter 7 presents two stages of the validation research: trialling and pretesting. Before completing the test specifications, a detailed description of the test and a trial version of the test tasks are compiled and test taker and rater feedback is collected through semi-structured interviews. After the qualitative data collection and analysis of the small-scale trial, I present the collection of quantitative data in the course of pretesting the proposed test tasks. As for pretesting, the recommended test development protocol was followed: the test paper was administered with a pretest population which was similar to the target population of the academic test; the result of the pretest was analysed using Classical Test Theory (CTT).

The findings of the qualitative and quantitative data analysis have twofold relevance. On the one hand, the analysis of the verbal protocols and the large-scale pretesting helped establish the validity of the writing tasks. On the other hand, they shed light on the issues of scoring validity I problematized in Chapter 8. In this chapter, I revisit the issues identified with rating in the previous stages and discuss the development of a genre and level specific checklist-based rating tool developed for ensuring the objective and unbiased nature of the rating procedure. Finally, the results of the thesis are summed up in the Conclusion chapter that indicates the suitability of rating on a checklist as a potential direction of the current research.

2. Research hypotheses and research questions

The aim of the present research is to build a validity argument about how the construct of the proposed writing tasks in an Academic test reflects the skills required in higher

education, and whether the results reflect reliable scores and unbiased marking. The dissertation presents the research-based test development process of the Euroexam English for Academic Purposes Test, with special attention to the context validity of Task 1 (*formal transactional email*) and an improved scoring validity of Task 2 (*discussion essay*). The research questions are as follows:

Research Question 1: Is transactional writing a valid task type for an EAP test?

Apart from discursive and argumentative writing, which appear both as authentic tasks in university education and in EAP tests, the main question concerns the validity of transactional writing in a test for Academic English. The research hypothesis implies that transactional writing is also part of students' repertoire. In addition to the professional side of academic life, university students are expected to arrange their studies, and develop and nurture issues in relation to administration and registration. Apart from meeting academic requirements, students are expected to meet the demands of formal communication regarding their studies. Based on this assumption, transactional writing is also part of the academic domain, therefore formal transactional text types are what students most often write in an academic context.

Chapter 6 of the dissertation investigates this question through empirical research and expert judgement. As part of the domain analysis a small-scale preliminary study was carried out with the following secondary research questions:

a) What are the most frequent written genres regarding communication between university undergraduates and members of staff?
b) Is formal written communication in English a part of university students' target language use (TLU)?
c) How important is the level of formality in TLU?

These questions aimed to disclose whether the proposed transactional writing task is suitable for an Academic exam using qualitative methods. The results of the study served as a basis for preliminary task design and the secondary research questions were addressed in the questionnaire used for expert judgement.

As regards scoring validity, the research questions are based on both quantitative and qualitative enquiry. The results of the verbal protocols and statistical analyses in Chapter 8 reveal the advantages of checklist-based assessment. The research hypothesis proposes that a task and level specific checklist-based assessment tool improves the

objectivity and reliability of the assessment of Task 2. The hypothesis is tested through the following research questions:

Research Question 2: Compared with a marking scale, can checklist-based assessment enhance

- the objective scoring of academic discussion essays and
- rater reliability?

The secondary research questions that are addressed in the course of the analysis using Classical Test Theory are as follows:

a) Is the reliability (Cronbach's alpha) of checklist scores high enough to fulfil accreditation requirements?
b) How do checklist items perform in terms of item difficulty and item quality?
c) Is the checklist capable of discriminating low and high performers?
d) Does checklist-based rating affect the success rate of the essay task?

Test scores are of particular importance for the different stakeholders of a test – universities, awarding bodies, test takers and raters. The main issue with the rating procedure of Euroexam is that test takers and raters have different perceptions of what counts as successful writing performance (Lukácsi, 2017). In addition to this, ratings may be subject to personal judgements and halo effect (Knoch, 2009), even "trained experienced raters have been shown to differ systematically in their interpretation of routinely-used scoring criteria" (Eckes, 2009, p. 5). Previous research at Euroexam International (Lukácsi, 2017; 2018, 2020) proved that a level and genre specific checklist enhances the objectivity and reliability of scoring a B2 level transactional writing task.

The verbal protocols with Euroexam raters in Chapter 7 were aimed to reveal how the raters approach the essay task during scoring. The verbal protocols also shed light on how the features they associate with a well-formed essay differ from each other. An additional qualitative enquiry in connection with scoring validity concerns Euroexam raters' ideas about the writing product:

Research Question 3: Can checklist-based marking increase the genre awareness of raters?

Based on the teacher verbal protocols and the rater think-aloud protocols in Chapter 7, a checklist-based rating tool is designed based on dichotomous statements and concept check

questions. Throughout the development of the level and task specific checklist, teachers' and raters' verbal protocols serve as a basis for qualitative analysis to design a checklist that may guide raters towards a common understanding of the genre of the essay.

The evidencing of objective and unbiased marking is one major requirement expected by international tertiary education institutions as the language requirements for university entrance have a gatekeeping function (Nagy, 2000). To make sure that applicants possess the skills based on their language certificates, it is important to design an assessment tool that does not only provide reliable scores, but also has a positive effect on the skills of the test takers. Language tests might have a positive effect on teaching practices and skills development, and have an "impact on the career or life chances of individual test takers" (Taylor, 2005, p. 154). Since the knowledge of English is regarded as a commodity (Cameron, 2000), it is of paramount importance that the test takers and university applicants are aware of the practices of the academic discourse (Weninger & Khan, 2013).

The marking procedure of writing tasks has always been an issue generating interest in language testing research. The scoring validity of the subjectively marked writing tasks – especially that of the academic discussion essay – is a key issue in this regard to examine. The tool most assessment related handbooks describe for the assessment of writing products is the rating scale (Alderson et al., 1995; Bachman & Palmer, 1996; McNamara, 1996; Shaw & Weir, 2007; Weigle, 2002, Weir, 2005). At the same time, the fallacy of subjective marking of learners' writing performance and the need for more objective, i.e. consistent assessment has been repeatedly raised by a number of publications (Eckes, 2009; Knoch, 2009; Knoch, 2011, Lukácsi, 2017; 2018; 2020; McNamara, 2000) as well as Chapter 9 of the *Common European Framework of Reference for Languages* (Council of Europe, 2001; 2018).

The use of an objective rating tool is expected to reduce differences among raters and increase their genre awareness. Apart from this immediate result, there is a predicted positive washback effect that will develop students' genre awareness and writing skills and also increase the probability of the correct perception of their writing results.

3. Methodology

The aim of the dissertation is to build a validity argument about how the construct of the proposed writing tasks in an academic test reflects the skills required in higher education,

and whether the results reflect reliable scores and unbiased marking. The method is built upon Weir's (2005) theoretical framework and the characteristics of test usefulness (Bachman & Palmer, 1996; 2010), and consider Read's (2015) validation stages, using a mixed-methods approach.

As discussed in the literature review, Bachman & Palmer (1996; 2010) introduced a model of test usefulness with six characteristics to consider in test development. In order to end up with a valid test, one must review the (a) reliability, (b) the construct validity, (c) the authenticity, (d) the interactivity, (e) the impact and (f) the practicality of the test tasks. Although Bachman and Palmer's work is still influential in test design, later Bachman (2005) pointed out that these categories are alone standing, and the relationship between them is not clearly defined. The only thing we can do is that we build "a convincing case that the decisions we make are defensible and supporting that case with credible evidence are the two components of the validation process" (Bachman, 2005, p. 5). When considering the six characteristics, we have to realise that it is impossible to achieve a high quality for all the characteristics, there are certain compromises we have to make, and instead of the 'perfect test', we have to focus on designing a 'good enough' test. Out of the six characteristics, the ones which are the most important for the purposes of EAP test design – to see how well the test tasks fit into the academic context and discourse (Chan, 2013) – are (a) reliability, (b) construct validity and (c) authenticity. It is important to design test tasks which provide comparable results in different administrations, measure what we want to measure and are representatives of target language use.

These three characteristics of Bachman and Palmer (1996) appear in Weir's (2005a) framework as the components of validity. As regards construct validity, Bachman and Palmer claim that it is essential that the construct is valid in a specific context. This idea was further developed by Weir (2005), who uses construct validity as an umbrella term and introduces new aspects of validity. In his framework, the construct is determined by the context, and authenticity appears as an integral part of context validity. In case of a writing task, context validity is about mapping the linguistic and content demands of a test task, and the demands of the real-life writing tasks in the target language, i.e. we have to see whether we are testing target language use in a specific context. He also introduces scoring validity, i.e. the validity of the rating procedure in which he integrated the notion of reliability (Shaw & Weir 2007; Weir, 2005).

The dissertation handles the question of validity for the two tasks in two different ways. The qualitative and quantitative parts of the research may be regarded as complementary, the method of mixing shows a sequential structure, i.e. the research shows an iterative structure in which results and conclusions of each stage are built in the design of the following stages (Creswell, 2009, p. 14). The four stages of validation and the processes I have specifically designed for the writing tasks of Euroexam Academic test are displayed in Table 1.

Table 1
Euroexam Academic Test Writing Tasks - Stages in Validation

Stage 1	**Stage 2**	**Stage 3**	**Stage 4**
Task 1 & Task 2	**Task 1 & Task 2**	**Task 1 & Task 2**	**Task 2**
Initial development	**Completion of test specifications and items**	**Pretesting test tasks**	**Establishing an improved scoring validity of checklist-based marking for essays**
Planning	Domain modelling and trialling	Evidence based analysis of test taker performance	Development of checklist items and CCQs
Domain analysis Preliminary investigation of the construct	Test taker characteristics	Student questionnaires	Verbal protocols
Expert judgement	Student and Rater interviews	Statistical analysis of pretest results	Rater and Candidate performance analysis

The steps of the stages of the development process were taken with regard to the writing construct of the Euroexam general C1 test. That is to say, both task types are examined in the first three stages, which are the standard stages of validation; however, the foci of the validation process are different for the two tasks. Stages 1 to 3 are the standard stages of validation research. Although I defined the focus and the main issues of my research in connection with the two tasks of the Academic test, I wanted the three standard stages to cover both tasks. The two tasks appear together in the initial development stage, trialling and pretesting stages. Stage 4 is an additional stage that was added to explore how an improved scoring validity of the essay task could be established.

The approach of the present research-based validation process uses construct validity as a feature to unify the arguments (Kane, 2013), the present validity argument is based both on theoretical and empirical evidence, where the different validities (context and scoring validity) are linked through their interaction (Shaw & Weir, 2007).

3. Results

In Stage 1, I conducted a small-scale preliminary study with the aim of defining the construct and to generate validity evidence for the context validity of transactional writing in the academic domain. In order to answer **Research Question 1** (Is transactional writing a valid task type for an EAP test?), I conducted student (N = 5) and staff (N = 6) interviews. I investigated what students write and how they communicate with university staff and aimed at establishing the validity of transactional writing as part of the academic discourse. Based on the answers to my secondary research questions of the preliminary investigation, my original hypothesis was confirmed: transactional writing is part of the academic domain. The mixed-methods research design was iterative in nature and involved triangulation to cross validate the findings of the elements of Stage 1 of the research. The results of the preliminary investigation were used to define the construct of the new Euroexam Academic test. The construct definition and the preliminary task design were the subject of the external expert judgement (N = 3). To enhance validity, the external experts were not provided with the empirical findings, but they reviewed the construct and the example tasks using a questionnaire. Using their knowledge and experience already available in the field, the experts found the proposed test tasks valid representations of the academic domain, and their comments helped me draft the specifications.

As for the validity evidence of discursive writing in the domain, I draw on literature review and used expert judgement to see what genres could be used as valid tasks in an EAP test. By using transactional writing and discursive writing for Task 1 and Task 2, respectively, it was possible to keep the writing construct as specified by the requirements of profile extension in the *Accreditation Manual* (Educational Authority, 2019).

After Stage 1, the aim of Stage 2 was to complete the test specifications and the example tasks that were to be pretested in Stage 3. Similarly to domain analysis, domain modelling in Stage 2 is also evidence focused. I used test taker performance and test taker and rater interviews to see whether the two proposed task types conform to the construct. At this point, it is important to highlight that the verbal protocols and the textual analysis in Stage 2 also served as the trialling of the exam tasks. When test taker and rater feedback was collected, I used the preliminary tasks which were designed at the end of Stage 1 for think aloud and immediate recall. The reason for this is twofold. On the one hand, the theoretical construct of the genres is not suitable for collecting user feedback. On

the other hand, through trialling example tasks, we may observe test taker characteristics and task characteristics; in other words, the validity of the rating process may be ensured.

The small-sample trial was conducted using qualitative methods: potential test taker interviews (N = 6) and Euroexam rater think-aloud protocols. The main aim of Stage 2 was to complete the test specifications, which was done based on the data gathered in the course of trialling. To gather data on the validity of the rating process, experienced accredited raters (N = 3) of Euroexam International took part in Stage 2. The three raters were given the same test taker scripts and were asked to assess them when using the accredited C1 level writing scale of Euroexam International through a think-aloud technique in Hungarian. I found that Euroexam raters varied in their scoring behaviour, their construct interpretations, and their severity. It seemed that the vague descriptors of the rating scale hindered the objectivity of the rating process and thus the reliability of the test scores.

In accordance with the cyclical nature of the design, at the end of Stage 2, I redesigned Task 1 based on the students' comments. In addition to this, a detailed task specification was completed for the Euroexam Academic Test. In Stage 3, the large-scale pretesting stage, I used the format and layout of the tasks as they were redesigned based on Stage 2.

The validation process in Stage 3 involved large-scale data collection and evidence-based analysis of test taker performance (N = 136). The aim of this stage was to check that test tasks work as intended so that the standard level of the Euroexam Academic test (C1) could be set, the relationship to the CEFR could be established and the validity of the test could be demonstrated. To ensure all this, I used a sample size (N = 136) for pretesting that allows statistical data analysis using Classical Test Theory (CTT). The aim of pretesting is to model the live administration of the test and to see how test takers and test tasks perform under exam circumstances. In addition to pretesting the tasks of the Academic Test, I used a questionnaire I designed to collect test taker personal data concerning their language learning background and self-assessment as well as test taker opinion of the form and content of the test.

The statistical analysis of test taker results proved that the Euroexam Academic Test is a valid measure of C1 level writing skills, however, raised further issues about the scoring validity of the two writing tasks. Due to the recursive process in my data analysis,

it became clear that the results of the earlier stages all lead in a certain direction: they highlighted raters' differences and flaws of the rating procedure. The results of Stage 2 and Stage 3, based on the verbal protocols, revealed that there was a discrepancy between the scores and the students' self-evaluation. The shortcomings of the C1 level accredited rating scale of Euroexam International were also exposed by the rater think aloud protocols. Chapter 7 revealed raters' ideas about the writing product and the rating scale, and revealed considerable rater bias.

Stage 4 of the research focused on the development of a checklist-based assessment tool following Lukácsi's (2017; 2018; 2020) research to increase the scoring validity and the reliability of the assessment of the essay task. The aim of this stage was to develop a task and level-specific checklist-based rating tool for the essay writing task that compensates for individual rater characteristics and rater effect. Stage 4 consisted of two main parts, document analysis (phase 1) and empirical research (phases 2-8). The empirical research was divided into two major steps: phases 2-6 focused on designing and developing the items of the checklist, and phases 7-8 aimed at exploring the relationship between scale-based and checklist-based scores.

Based on the findings in Stage 4 of the research-based validation process, it can clearly be stated that the use of the proposed checklist-based assessment tool improves the scoring validity of the essay task of the Euroexam Academic Test. The results of the checklist development process and the large-scale field test support the original research hypothesis. The methodology of the validation research, i.e. mixed-methods research is reflected in the research questions I formulated at the beginning of the research. Research Question 2 targeted the quantitative part of the research, whereas Research Question 3 was used as a qualitative cross-validation. Following the two main research questions, and the Secondary research questions of the second research question, the checklist development project led to the subsequent conclusions:

Research Question 2: Compared with a marking scale, can checklist-based assessment enhance

- the objective scoring of academic discussion essays and
- rater reliability?

As the dichotomous items and the concept check questions leave less chance to rater bias, checklist-based assessment increases rater objectivity. Based on the figures, checklist-based assessment increases the scoring validity of the test. The higher level of inter-rater

reliability was demonstrated through various statistical analyses (exact agreement, ICC, and Krippendorff's alpha).

The secondary research questions targeted specific statistics in the course of the development project. Objectively, these figures suggest increased scoring validity for the discussion essay on their own; furthermore, the values fulfil the Hungarian accreditation requirements (Educational Authority, 2019).

a) Is the reliability (Cronbach's alpha) of checklist scores high enough to fulfil accreditation requirements?
   The reliability of the scores of based on the final 30-item checklist (α = .90) fulfils the accreditation requirement of $\alpha \geq .75$.
b) How do checklist items perform in terms of item difficulty and item quality?
   Concerning item difficulty and item quality the item level statistics conform to the specifications of the *Accreditation Manual*, namely that more than 80% of the p-values and 90% of the discrimination indices (Ebel's D) fall within the acceptable range of .70 ≥ p-value ≥ .30; Ebel's D ≥ .30.
c) Is the checklist capable of discriminating low and high performers?
   The high values for Ebel's D indicate that the checklist is capable of discriminating high and low performers. This is also clearly discernible on the frequency distribution chart (Figure 21), in which we can observe a broader score range without the presence of a central tendency.
d) Does checklist-based rating affect the success rate of the essay task?
   Although the scores are spread out, the success rate of the essay task calculated with a paired samples t-test using scale-based and checklist-based results is not different, therefore checklist-based assessment is not more severe that scale-based assessment, and it does not affect the standard.

As for **Research Question 3** (Can checklist-based marking increase the genre awareness of raters?), the answer may be given based on the results of the qualitative data analysis of Stage 4. The dichotomous items of the checklist and the concept check questions increased the genre awareness of raters. The majority of the feelings the participants expressed in the course of item development are about increased objectivity and a positive attitude towards a tool that gives clear-cut criteria. They all seemed to be happy to follow these instead of relying on "gut feelings".

The results of the checklist development project may lead to the conclusion that it is possible to minimise rater bias, reduce the strong central tendency in rating (Eckes et al., 2016), and direct raters toward a common understanding of assessment and genre criteria. Furthermore, an analytical scale that focuses on directly observable phenomena may enhance teacher's feedback practices and thus increase positive washback. Based on the results of the checklist development project, we can claim that the checklist-based rating tool has a number of advantages. All things considered, the research design could be adapted to develop a similar rating tool for the transactional writing task, as well as all for the genres that appear in the writing paper of the C1 level Euroexam General English Test.

Conclusion

The dissertation aimed to present the research-based validation process of the writing tasks of the English for Academic Purposes (EAP) test of Euroexam International and the development of a checklist-based rating tool for the assessment of discussion essays within the academic domain. The research project was motivated by the endeavour of Euroexam International to design and implement a locally developed EAP test, and by my interest in the assessment of writing skills and the possible ways of increasing the objectivity of the rating process. The most important findings of the dissertation concern the validity of the writing tasks of the locally developed EAP test of Euroexam International. Based on the results of the 4-stage research-based validation process, it has been confirmed that the test tasks are valid measures of English language skills within the academic domain. A further contribution of the research is the development and validation of a checklist-based rating tool, the use of which results in an increased scoring validity and a more reliable rating for the discussion essay task.

References

Alderson, J. C., & Clapham, C. (1995). Assessing student performance in the ESL classroom. *TESOL Quarterly, 29*(1), 184–187.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Cameron, D. (2000). Styling the worker: Gender and the commodification of language in the globalized service economy. *Journal of Sociolinguistics, 4*(3), 323–347.

Chan, S. H. C. (2013). *Establishing the validity of reading into writing test tasks for the UK academic context*. PhD thesis. University of Bedfordshire. Retrieved on 16 June 2020, from: http://uobrep.openrepository.com/uobrep/handle/10547/312629

Council of Europe (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press. Retrieved on 20 June 2020, from https://rm.coe.int/1680459f97

Council of Europe (2018). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume with new descriptors*. Strasbourg: Language Policy Division. Retrieved on 20 June 2020, from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989

Creswell, J. W. (2009). *Research Design. Qualitative, Quantitative and Mixed Methods Approaches*. Thousand Oaks, CA: Sage Publications.

Eckes T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the common European framework of reference for languages*: *Learning, teaching, assessment (Section H)* (pp. 1–52). Strasbourg, France: Council of Europe/Language Policy Division.

Eckes, T., Müller-Karabil, A., & Zimmermann, S. (2016). Assessing writing. In: D. Tsagari, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 147–164). Boston: De Gruyter Mouton.

Educational Authority (2019). *Accreditation manual 2019*. Retrieved on 16 June 2020, from https://nyak.oh.gov.hu/nyat/doc/ak2019/ak2019.htm

Educational Authority (2020). *Basic definitions*. Retrieved on 16 June 2020, from: https://nyak.oh.gov.hu/doc/alapfogalmak-eng.asp

Fűköh, B. (2019a, May). *Research-based EAP test development: local needs and opportunities on an international context.* Paper presented at the 14th UPRT Conference, University of Pécs, Pécs, Hungary.

Fűköh, B. (2019b.) Kutatáson alapuló tesztfejlesztés – írásfeladatok egy angol tudományos szaknyelvi vizsga számára. In: (Besznyák R. (Ed.) *Porta Lingua – 2019. Interdiszciplináris megközelítések a szaknyelvoktatásban és –kutatásban* (pp. 271–288). Budapest: Szaknyelvoktatók és -Kutatók Országos Egyesülete. http://szokoe.hu/kiadvanyok/porta-lingua-2019

Government Decree 137/2008. (V. 16.) Korm. rendelet az idegennyelvtudást igazoló államilag elismert nyelvvizsgáztatás rendjéről és nyelvvizsga bizonyítványokról. Retrieved on 16 June 2020, from: http://njt.hu/cgi_bin/njt_doc.cgi?docid=119127

IELTS (2018). *IELTS test format.* Retrieved on 16 June 2020, from: https://www.ielts.org/about-the-test/test-format

Kane, M. T. (2013). The argument-based approach to validation. *School Psychology Review*, *42*(4), 448–457.

Knoch, U. (2009). *Diagnostic assessment of writing: The development and validation of a rating scale*. Frankfurt: Peter Lang.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing*, *16*, 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Lukácsi, Z. (2017, May). *Developing a level-specific checklist for assessing writing*. Paper presented at the 14th EALTA Conference, Ciep, Sèvres, France.

Lukácsi, Z. (2018). Írásművek lista alapú értékelése – avagy hogyan mérjünk a nyelvvizsgán. *NyelvVilág, XXI,* 7–23.

Lukácsi, Z. (2020). Developing a level-specific checklist for assessing EFL writing. Language Testing. https://doi.org/10.1177/0265532220916703

McNamara, T. F. (1996). *Measuring second language performance*. New York, NY: Longman.

McNamara, T. (2000). *Language testing.* Oxford: Oxford University Press.

Nagy, P. (2000). The three roles of assessment: Gatekeeping, accountability, and instructional diagnosis. *Canadian Journal of Education*, *25*(4), 262–279.

O'Sullivan, B., & Dunlea, J. (2015). *Aptis general technical manual*. London: British Council.

Pearson Academic (2017). *Test format*. Retrieved on 16 June 2020, from https://pearsonpte.com/the-test/format/

Read, J. (2015). *Assessing English proficiency*. Palgrave: Macmillan.

Shaw, S., & Weir, C. J. (2007). *Examining writing: Research and practice in assessing second language writing*, Studies in Language Testing (Vol. 26). Cambridge: UCLES/Cambridge University Press.

Taylor, L. (2005). Washback and impact. *ELT Journal, 59*(2), 154–155.

TOEFL iBT (2018). *Test content.* Retrieved on 16 June 2020, from https://www.ets.org/toefl/ibt/about/content/

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: an evidence-based approach*. Basingstoke: Palgrave Macmillan.

Weninger, C., & Khan, K. H.-Y. (2013). (Critical) language awareness in business communication. *English for Specific Purposes, 32,* 59–71.

Relevant publications

Fűköh, B. (2016a). Developing Writing Skills of Students of Business English. *NyelvVilág, XX,* 7–18.

Fűköh, B. (2016b). Nyelvismeret, nyelvtanulási motiváció és második idegen nyelv: a nyelvi és motivációs felmérés néhány tanulsága a BGE turizmus–vendéglátás szakos I. évfolyamos hallgatók egy csoportjában. In: T. Váradi (Ed.) *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből 2016*: *X. Alkalmazott Nyelvészeti Doktoranduszkonferencia* (pp. 19–30). Budapest: MTA Nyelvtudományi Intézet.

Fűköh, B. (2018). Student interviews in establishing the context validity of an EAP writing task. In: (R. Besznyák (Ed.) *Porta Lingua - 2018. Tudásmegosztás, értékközvetítés, digitalizáció– trendek a szaknyelvoktatásban és -kutatásban* (pp. 301–309). Budapest: Szaknyelvoktatók és -Kutatók Országos Egyesülete. http://szokoe.hu/kiadvanyok/porta-lingua-2018

Fűköh, B. (2019). Kutatáson alapuló tesztfejlesztés – írásfeladatok egy angol tudományos szaknyelvi vizsga számára. In: R. Besznyák (Ed.) *Porta Lingua – 2019. Interdiszciplináris megközelítések a szaknyelvoktatásban és –kutatásban* (pp. 271–288). Budapest: Szaknyelvoktatók és -Kutatók Országos Egyesülete. http://szokoe.hu/kiadvanyok/porta-lingua-2019

Fűköh, B. (2017). Language learning gains and motivation of L2 and L3 learners of ESP. In: M. Lehmann, R. Lugossy, M. Nikolov, & G. Szabó (Eds.) *UPRT 2017: Empirical Studies in English Applied Linguistics* (pp. 177–188). Pécs: Lingua Franca Csoport.

Lukácsi, Z., & Fűköh, B. (2018). Vizsgafejlesztés a magyar felsőoktatásban: Egy egyetemi szaknyelvi vizsga kidolgozása és fogadtatása. *Új Pedagógiai Szemle, 68*(9-10), 42–63.

Lukácsi, Z., & Fűköh, B. (2020). A tantervi hatékonyság vizsgálata az egyetemi szaknyelvoktatásban. *Modern Nyelvoktatás, 26*(1-2), 28–43.

Conference presentations

Fűköh, B. (2018, September). Establishing the Context and Scoring Validity of an English for Academic Purposes test. In: J. Rambousek, I. Schöfrová, & J. Chamonikolasová (Eds.) *14th ESSE Conference Abstracts* (p. 138). Brno: Masaryk University.

Fűköh, B. (2018, September). Validity Arguments for English for Academic Purposes Test Tasks. In: J. Rambousek, I. Schöfrová, & J. Chamonikolasová (Eds.) *14th ESSE Conference Abstracts* (p. 133). Brno: Masaryk University.

Fűköh, B. (2019, May). *Research-based EAP test development: local needs and opportunities on an international context.* Paper presented at the 14th UPRT Empirical Studies in Applied Linguistics Conference, University of Pécs, Pécs, Hungary.

Lukácsi, Z., & Fűköh, B. (2018, June). *A truly special examination: A case study of a local exam development project*. Paper presented at the 13th UZRT Empirical Studies in Applied Linguistics Conference, University of Zagreb, Zagreb, Croatia.

Lukácsi, Z., & Fűköh, B. (2018, November). *Assessing academic English among Central European students for UK University admissions*. Poster presented at the Language Testing Forum 2018 Conference, University of Bedfordshire, Luton, UK.

Lukácsi, Z., & Fűköh, B. (2019, May). *Assessing academic English with a localised test for international admissions purposes*. Paper presented at the 16th EALTA Conference, SIG Assessment of Writing / Assessment for Academic Purposes, University of Dublin, Ireland.