Tests as a Technique for an ESP Coursebook Evaluation

Eva Ellederová

Department of Foreign Languages, Faculty of Electrical Engineering and Communication

Brno University of Technology, Czech Republic

Biodata

Eva Ellederová received her Ph.D. in foreign language pedagogy from Masaryk University in Brno, the Czech Republic. She graduated from the Faculty of Arts, Masaryk University with a degree in English Language and Literature, but her first degree in Process Engineering was earned from Brno University of Technology. She has been teaching English since 1993. At present, she works as a director of the English Language Section at the Department of Foreign Languages of the Faculty of Electrical Engineering and Communication at Brno University of Technology where she teaches English for Information Technology, Business English, and Practical English courses. She currently focuses on design-based research of ESP coursebooks and analysis of information technology students' speaking skills.

Email: elleva@seznam.cz

Abstract

Design-based research evolved near the beginning of the 21st century as a practical research methodology that could effectively bridge the gap between research and practice in formal education, as it aims at both developing theories about domain-specific learning and the means designed to support that learning. Since ESP teachers often design their own learning

materials, this kind of research enables them not only to evaluate their quality and gradually improve them but also to produce design principles. Accordingly, the aim of my ESP coursebook design-based research is to evaluate the quality of a coursebook pilot version I developed for the students from the Faculty of Information Technology at Brno University of Technology in the Czech Republic, and consequently formulate necessary modifications based on design principles. I first frame the basic concept of design-based research and ESP testing. Then, I describe the methodology of the ESP coursebook evaluation. Results indicate the requirements for the coursebook redesign, which involve adding more tasks for the acquisition of linguistic means for expressing different language functions, including more material for recycling and reinforcement focused mainly on vocabulary practice and increasing the level of difficulty of listening passages. Based on these results, I discuss the ESP coursebook design principles related to multi-skill tasks, and professional vocabulary and language functions acquisition.

Key words: ESP coursebook evaluation, design-based research, test specification, design principles

1. Introduction

Despite a wide variety of ESP coursebooks available on the market, it is still rare to find those meeting both particular course requirements and students' needs. Some texts and topics in the published coursebooks seem to be irrelevant and outdated for particular ESP courses or study programmes (e.g. information technology, chemistry, electrical, mechanical engineering and civil engineering), and it is often quite difficult to cover a one-semester course with the exact number of units in published coursebooks (Mol & Tin, 2008; Vičič, 2011; Barnard & Zemach, 2014; Danaye & Haghigi, 2014; Ellederová, 2020). For this reason, ESP teachers must either adapt existing learning materials or design their own materials. Therefore, the need arises to design a coursebook that is tailored for the ESP course. The purpose of my ESP coursebook design-based research (DBR) is to gradually develop the coursebook based on its iterative evaluation in the classroom environment.

One of the methods of combined evaluation (evaluation that uses different kinds of evaluation tools such as questionnaires and tests) of an ESP coursebook are tests used to verify students' knowledge and skills acquired after using the coursebook. Besides evaluation of an ESP

coursebook by means of a checklist or a survey conducted among teachers and students (e.g. Mol & Tin, 2008; Jebahi, 2009; Zangani, 2009; Razmjoo & Raissi, 2010; Habtoor, 2012; Barnard & Zemach, 2014; Danaye & Haghigi, 2014; Ellederová, 2018), a coursebook quality can also be assessed through summative performance evaluation (Alderson, 1988). This type of evaluation might take place at the beginning and at the end of a course to determine whether the coursebook has been effective. Pre-testing and post-testing students' knowledge and skills also serves as a part of quality control, which helps improve the coursebook by changing some of its aspects.

In this study, I describe the two phases of an ESP coursebook DBR focused on testing as a method for evaluating knowledge and skills students are expected to have acquired after using the pilot version of the coursebook *English for Information Technology* (Ellederová, 2016). I designed the coursebook for the course English for Information Technology (IT) as part of the bachelor's degree study programme Information Technology at the Faculty of Information Technology, Brno University of Technology (BUT), in the Czech Republic. The evaluation of the coursebook quality by means of testing should lead to the improvement of its pilot version and to the empirically grounded design principles for this particular ESP coursebook, with implications for ESP coursebooks in similar contexts. To this end, the following sections frame the concept of DBR and ESP testing, describe the development and verification of tests, present and analyse the collected data and discuss research findings including the impact on the coursebook redesign and the preliminary design principles.

2. Literature Review: Concept of Design-Based Research and ESP Testing

Researchers from various domains of education have discussed the need for a research design to adequately address problems in educational practice and develop designs, which will be tested through pragmatic experiments and grounded in learning sciences. Design-Based Research Collective (2003) notes that "educational research is often divorced from the problems and issues of everyday practice – a split that creates a need for new research approaches that speak directly to problems of practice and that lead to the development of 'usable knowledge'" (p. 5). DBR was introduced with the expectation that "researchers would systematically adjust various aspects of the designed context so that each adjustment served as a type of experimentation that allowed the researchers to test and generate theory in naturalistic contexts" (Barab & Squire, 2004, p. 3). Gradually DBR has become an emerging

paradigm for the study of teaching and learning in context through the methodical design and study of instructional strategies and tools. Van der Akker (2006) emphasizes that "the merit of a design is measured, in part, by its practicality for users in real contexts" and adds that "the design is (at least partly) based upon theoretical propositions, and field testing of the design contributes to theory building" (p. 5). Greater attention paid to the processes of results selection and analysis relevant to practitioners, development of models of teaching and learning based on them, and systematic dissemination of these innovations increase the probability of the practitioners' use of the outcome of DBR to improve the quality of teaching and learning.

DBR, therefore, is a systematic, flexible methodology focusing on the improvement of educational practice by means of an iterative analysis, design, development and implementation of educational interventions, based on the collaboration of researchers and practitioners in naturalistic context, which results in the production of design principles and new theories (Barab & Squire, 2004; Plomp & Nieveen, 2013; Reeves & Amiel, 2008; Van der Akker, 2006). DBR is situated in a real educational context and it focuses on the design and testing of a significant intervention (e.g. an educational program, learning environment, teaching-learning method, learning material), uses mixed methods of data collection, involves multiple iterations, evolution of design principles, a collaborative partnership between researchers and practitioners, and has a practical impact on practice by means of attempting to find a solution to a complex educational problem and making practitioners reflect upon the results of their research (Van der Akker, 2006; Bakker & Van Eerde, 2013; Plomp & Nieveen, 2013). One of the distinctive characteristics of DBR is the twofold yield (see Table 1), namely, research-based interventions as well as knowledge about them, or theories based on them (Plomp & Nieveen, 2013). The challenge for DBR is to "capture and make explicit the implicit decisions associated with a design process, and to transform them into guidelines for addressing educational problems" (Plomp & Nieveen, 2013, p. 22). Its aim to contribute to the body of scientific knowledge (for development studies) or to generate or validate theories (for validation studies) distinguishes DBR from just systematic educational design processes which aim solely at designing educational materials through iterative cycles of testing and improving prototypes without interweaving the design with testing and theory development. Although DBR is related to and includes action research, the essential difference is that action research is not aimed at generating design principles and providing an empirically grounded

theory on how the research-based intervention works (Van der Akker, 2006; Bakker & Van Eerde, 2013; Plomp & Nieveen, 2013).

Table 1: Twofold Yield of DBR*.

Type of study	Research goal	Twofold yield
Development	Development of intervention	1) developing a research-based intervention as solution to complex problem, *and* 2) constructing (re-usable) design principles
Validation	Theory development and/or validation	1) designing learning environments *with the purpose* 2) to develop and validate theories about learning, learning environments, or to validation design principles

* Adapted from Plomp and Nieveen (2013, p. 23).

ESP testing relates to the area of language testing where the content of the test and the testing methodology are based on the analysis of a situation in which the specific language is used. While the objective of general English tests is defined more generally, ESP tests are usually defined more narrowly and a test designer must consider the specific purposes such as academic, occupational, technical, scientific, and medical. For this reason, Douglas (2013) emphasizes the importance of the three characteristic aspects of language for specific purposes (LSP) that influence the design of tests: (i) language use varies with the context; (ii) LSP is precise; (iii) there is an interaction between LSP and specific purpose background knowledge.

A significant difference between ESP assessment and assessment in other areas of language learning is the relationship between ESP and field knowledge. Douglas (2013) points out that

> over the years, practitioners have gradually come to the realization that language knowledge and background knowledge are very difficult to distinguish in practice and that, although specific purpose testers are not in the business of assessing professional, vocational, or academic competence in specific purpose fields, such competence is inextricably linked to language performance in those fields (p. 369).

According to Douglas (2013), two theoretical questions influence the course of ESP testing: whether there is indeed a concept of defined capabilities in ESP and how we should deduce criteria for assessing ESP performance. Davies (2001) argues that what often characterizes ESP test tasks is content rather than language itself, and therefore ESP testing should be based on its practical need and pragmatic efficiency since it

> cannot be about testing for subject specific knowledge. It must be about testing for the ability/abilities to manipulate language functions appropriately in a wide variety of ways. This might mean no distinction between a general proficiency test and an LSP test. […] No doubt for face validity reasons, the stimuli in such tests will be field related. (p. 143)

Based on Davis's assertion, the main reason for developing ESP tests is to ensure face validity and the equivalent level of test difficulty and expertise. The weakness of Davis's concept of ESP tests is the over-emphasis on exclusively language elements of ESP communication. The performance of ESP also depends on the relevant non-linguistic knowledge (i.e. knowledge of actual entities, states, relationships, regularities and generally assumed facts), which is inseparably linked with language knowledge. According to Jacoby and McNamara (1999),

> to assess special-purpose performance with general linguistic criteria, however, seems oddly out of synch with long-held fundamental positions in special-purpose language pedagogy and research, e.g., that special-purpose language is only a means to the acquisition of nonlinguistic knowledge and skills; that using the traditional four linguistic skills to delineate special-purpose performance is inadequate to capture real-world communicative cultures and activities; and that special-purpose performance is by definition task-related, context-related, specific, and local. (p. 234)

Therefore, when developing an ESP test, a test designer must assume that students' needs should not only include language skills, but also discipline knowledge appropriate to the communication context in which they study and work. This knowledge affects the understanding of the text and the overall performance of the students in the test. Alderson (1988) remarks that besides the use of authentic texts, "authentic purposes for language use" (p. 90) are important for language testing. The theoretical framework of ESP testing must be extended to include not only well-defined linguistic characteristics, but also characteristics of the context of interest of those who are tested. Considering the different contexts, performance will, for example, vary between technical and humanities students, doctors and air traffic controllers, hotel receptionists and supermarket vendors. Besides, doctors use a different language when talking to colleagues and patients. Hymes (1974) lists the contextual factors that influence how we use and accept language. These include the environment and

situation, participants (speaker and listener), ends (purpose of the event and individual goals of the participants), sequence of speech acts (their organization and content), tone of speech, instrumentalities (language, dialect, variety and channel), standard of interaction, and genre or type of event.

The materials used for the test design must involve students in tasks in which both language skills and field knowledge interact with the test content in a way similar to the situation in which the target language is used. Douglas (2000) emphasizes that test tasks "must be authentic for [the test] to represent a specific purpose field in any measurable way" (p. 6), which implies that ESP-specific testing must be done using field-specific content when designing tasks. Bachman and Palmer (1996), too, state that the design of any test must be based on the specific purpose, the group of test takers, and the target language use (TLU) domain, which is defined as a "set of specific language use tasks that the test taker is likely to encounter outside of the test itself" (p. 44). In practice, this means that, for example, the listening test tasks can be based on an online tutorial on graphics software or an interview about new trends in video games design for students of computer science. The test tasks must be of a similar nature to those performed by students in computer science courses at university or in a real work environment. To interpret the students' performance in the test as proof of language ability when using ESP, we must engage the test takers in the tasks that authentically represent the situation.

Another important factor in ESP testing is the washback effect that refers to negative or positive impact that the test has on teaching and learning (Dudley-Evans & St John, 1998). In ESP testing, this effect plays a crucial role, and its analysis enables to determine if the test reveals deficiencies and helps students acquire areas of subject matter which they still do not understand properly. The washback effect analysis might also reveal that students may not improve their reading comprehension significantly, but the test has a significant effect on students' attitudes and the way and the content of teaching (Jafarabadi et al., 2014). Therefore, an ESP test based on a students' language needs analysis will be more beneficial .

According to Pearson (1988), a good test should cover all areas of a syllabus, promote the use of beneficial teaching and learning processes and should be directly applicable as a learning activity. Similarly, it is recommended to use texts and activities in an ESP coursebook as the basis for the design of tests taken at the end of the ESP course.

3. Research Objectives and Research Questions

Despite the significance of DBR outlined above, there are only a few empirical studies that used pre- and post-testing as combined evaluation of learning materials. Also, those that exist were conducted at lower levels of education. For example, in the Czech Republic, Trna (2011) tested knowledge and skills of 80 elementary school learners after and before using physics worksheets focused on the topic of application of physics in everyday life. Another study conducted in the United States by Baumann et al. (2013) dealt with English language vocabulary acquisition of 606 elementary school learners. They pre- and post-tested learners' vocabulary during the development MCVIP (Multifaceted, Comprehensive Vocabulary Instruction Program) aimed at vocabulary learning. Both studies proved that pre-testing and post-testing learners' knowledge enables to reveal weak and strong aspects of the evaluated intervention that should be further modified and adapted. Unfortunately, these studies do not focus on ESP coursebooks, they do not provide the description of test construction and verification at the university level. This research is expected to fill this gap. Toward this end, the objectives of my research were:

1. to collect information about the coursebook quality by means of testing students' knowledge and skills before and after using the coursebook, and

2. to suggest the coursebook redesign according to the test results, and consequently produce preliminary design principles.

With these in mind I framed the following research questions:

1. Are the students' results (i.e. average test score) in the pre-test and the post-test different?

2. Are the students' pre- and post-test results in the subtest Use of English different?

3. Are the students' pre- and post-test results in the subtest Reading different?

4. Are the students' pre- and post-test results in the subtest Listening different?

I believe the answers to these questions enabled me to determine a) the quality of the evaluated ESP coursebook based on testing students' knowledge and skills, and b) the changes that needed to be made in the design of the ESP coursebook pilot version.

Since the course English for IT focuses on the development of students' skills and the acquisition of specialized vocabulary and language functions, I framed the following sub-questions:

It is important to note here that the students learn and develop writing skills in the course English for Europe (provided by the Department of Foreign Languages at BUT) focused on academic writing in an IT context. Therefore, the course English for IT as well as the coursebook aim at the development of three skills: reading, listening, and speaking. The quality of speaking tasks and activities in the coursebook was evaluated by teachers and students in another stage of the research (see Ellederová, 2018) which is not described in this study.

4 Methodology

4.1 Research Design

The research design was divided into one preparation phase and three realization phases. The preparation phase, which lasted from September 2017 to June 2018, focused on:

1. gaining an insight into the current state of knowledge of DBR of ESP learning materials,
2. construction of the first version of tests,
3. piloting and modification of tests to verify their equivalence and reliability, and
4. construction of the final version of tests.

The first realization phase, which took place in winter and summer semester in 2018/2019,

involved the following stages:

1. pre-testing of students at the beginning of the course,
2. implementation of the coursebook in lessons,
3. post-testing of students after they finish the course, and
4. data analysis and interpretation.

The coursebook redesign follows the above-mentioned stages as well as the stages including teachers' and students' evaluation of the coursebook described in Ellederová (2018). The second realization phase involves repeated implementation of the coursebook, data collection and analysis, results evaluation and discussion. The aim of this phase is the second data analysis and interpretation. The third realization phase focuses on the production of substantive and procedural design principles. Its aim is to characterize the optimal coursebook design, formulate design principles and to draw up recommendations designed to improve educational practice.

4.2 Pilot Version of the Coursebook

The methodological and pedagogical concept of the coursebook *English for Information Technology* is based on a combination of the following syllabi: topic-based syllabus (titles of individual units are based on topics from the field of information technology), skill-based syllabus (coursebook units are organized around the individual skills), functional syllabus (units focus on the acquisition of different language functions related to the particular topic), lexical syllabus (lexical items are divided into groups according to topics including the wordlist at the end of each unit), competency-based syllabus (tasks and activities in the coursebook focus on the development of competencies that students have to master in a given situation (e.g. a job interview focused on a particular information technology career, hardware and software troubleshooting, persuading potential customers about the quality of software or hardware), and task-based syllabus (coursebook units include tasks for solving various problems that students have to solve through communication in the target language).

The coursebook is aimed at the intermediate level learners who study information and communication technology at universities and wish to pursue their careers in this field. Its aim

is to equip the university students with both receptive and productive skills in professional English language at the level B2 of the *Common European Framework for Languages* (CEFR) focused on information and communication technology. The coursebook enables students to acquire professional vocabulary, linguistic means for expressing different language functions and to develop language skills necessary for both active participation in the seminars, lectures, conferences, and effective communication with colleagues, business partners and institutions in the competitive international environment of the information and communication technology sector.

The pilot version of the coursebook consists of fourteen main units covering a wide range of topics dealing with information and communication technology, one review unit, answer key and audio transcript. The units focus on the current development and careers in information technology, hardware, software, networks, the Internet safety, and the computer and the Internet history. Each unit consists of a lead-in activity, main topic text, vocabulary practice, reading, listening, speaking tasks and language functions, such as predicting, classifying, giving instructions, persuading, describing features and processes. All tasks correspond to the Cambridge English exams format; they include multiple matching, gap filling, multiple-choice cloze, multiple choice, sentence completion and true/false tasks. The English-Czech wordlist with phonetic transcription of the specialized terminology accompanies each unit. Most tasks are based on the communicative approach in language learning, but the elements of the Presentation-Practice-Production method and task-based approach are implemented in the coursebook as well.

4.3 Research Participants

The participants were the first-year students of the bachelor's degree study program Information Technology at the Faculty of Information Technology at BUT. The students' native languages were predominantly Czech and Slovak. Five students were Russian. During the first realization phase, 92 students participated in pre-testing and post-testing. The students (divided into five groups) attended one-semester course English for IT (two hours every week) which I supervise and teach. The prerequisite for enrolling on the course English for IT is successful completion of the B1 level course of academic English.

54% of the students studied English from 11 to 15 years and 73% graduated from the secondary school with the state school-leaving exam in English whose CEFR level is B1; therefore, they might be considered as a homogenous group with the same entry language level corresponding to the course requirements. Besides, three of the students had the First Certificate in English and two had the Certificate in Advanced English. One student passed the Pearson LCCI International Qualifications exam (level B1) and another passed the International Baccalaureate exam (Higher Level).

4.4 Construction and Verification of Tests in English for IT

Since there are no standardised ESP tests focused on information technology, it was necessary to construct my own tests. I designed a pre-test and a post-test in accordance with the evaluated coursebook *English for Information Technology*. They should verify students' knowledge and macro- and micro-skills in ESP focused on IT. First, I considered the following important factors related to the test construction (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2001; Alderson, Clapham, & Wall, 1995; Bachman & Palmer, 1996; Chráska, 1999; Weir, 2005): determining the purpose, type and objective of the test; characteristic of language input; characteristic of target language use (TLU) domain; characteristic of test content; range, type and a number of tasks in the test; total score and cut-off score; time allocated for the test.

The designed tests are criterion referenced tests that combine elements of a proficiency test and they are used as final tests for the course English for IT since their content and form are based on the content and form of tasks in the evaluated coursebook *English for Information Technology*. Criterion-referenced tests are designed to measure students' performance against a fixed set of predetermined criteria or learning standards, i.e. concise, written descriptions of what students are expected to know and be able to do at a specific stage of their education, and they are used to evaluate whether students have learned a specific body of knowledge or acquired a specific skill set (Chráska, 1999, Schindler et al., 2006). Proficiency tests are designed to show whether "students have sufficient ability to be able to use a language in some specific area such as medicine, tourism, or academic study" (Alderson et al., 1995, p. 293). Content validity of both versions of the test (pre-test and post-test) were revised and commented on by teachers from the Department of Foreign Languages of the Faculty of Electrical Engineering and Communication (FEEC) at BUT.

In order to test the level of acquired knowledge and skills, I included three subtests in the test: Use of English, Reading and Listening. The reason for division of the test into three subtests was not only because the tests designed at the Department of Foreign Languages have the same or very similar format as standardized Cambridge English tests, but also because the aim was to determine which tasks were the most difficult/easiest for students, and consequently optimize the particular parts and tasks in the coursebook.

I based the topics and genres of texts and recordings used in the test on the evaluated coursebook *English for Information Technology*, scientific literature and multimedia specialised in IT, such as programming and computer science books and textbooks, scientific journals, magazines and web portals. I selected the language level, task types and characteristic in accordance with the coursebook, the CEFR and *Global Engineers Language Skills* (*GELS*) *Framework*. *GELS Framework* is the CEFR adaptation designed for university students of engineering study programmes. GELS project is a common initiative between the University of Cambridge, KTH Royal Institute of Technology of Stockholm and a French research laboratory (Institut Mines-Telecom – Didalang). The objective of this project is to enhance our future engineers' language skills in order to prepare them for the increasingly challenging demands of a globalised market (for more details see Rinder, Geslin, & Tual, 2016).

Chráska (1999) points out that "the crucial issue concerning test construction is a selection of subject matter content the [student] has to master" (p. 16); therefore, each tested aspect should be covered by a relatively large number of tasks. In the test, I included three tasks (40 items in total) for testing the level of vocabulary and grammar acquisition, three tasks (18 items) for testing reading skills acquisition, and two tasks (14 items) for testing listening skills. I determined the relative weights of points for each item and the time allocated for the whole test according to the standard for testing established by the Department of Foreign Languages at FEEC. The subtest Use of English focuses on both professional vocabulary and language functions; hence, the total number of points is highest here (see Appendix A).

I set a success rate of ≥ 70 % (i.e. ≥ 50 points) as a criterion. This criterion is adopted as a standard at Cambridge English exams specialised in ESP (see BEC Vantage) as well as at the final test in the course English for IT. I also set the same success rate for each subtest, i.e. the success rate for the subtest Use of English was ≥ 28 points, for Reading ≥ 13 points and for Listening ≥ 10 points. To pass the test, a student must pass each subtest.

A detailed specification of the test including language input characteristic, TLU domain, total number of points, cut-off score, task types and their description, genre and topics is shown in Appendix A (includes specification of both test forms, i.e. the pre-test and the post-test).

4.5 Verifying the Equivalence of Test Forms

According to Alderson et al. (1995) and Weir (2005), two or more forms of a test are interchangeable if they have the same objective and purpose, the same instructions, response types and number of items, and if they are based on the same content. Alderson et al. (1995) also emphasise that equivalent forms of a test should "measure the same language skills and that they correlate highly with one another" (p. 97). Similarly, Jackson (2009) recommends that the alternate forms of a test should have "the same number of items, the items should be of the same difficulty level, and instructions, time limits, examples, and format should all be equal – often difficult if not impossible to accomplish" (p. 68). Alternate forms of a test are two or more forms of a test that are interchangeable because they measure the same constructs in the same ways, are intended for the same purposes, and are administered using the same directions. It is a generic term used to refer to any of the following three categories parallel, equivalent and comparable (American Educational Research Association, American Psychological Association & National Council on Measurement in Education, 2001; Weir, 2005).

One means of ensuring equivalency of different test forms is to determine alternate-forms reliability – using alternate forms of the test and correlating the performance of individuals on the two different forms. Jackson (2009) uses the term alternate-forms reliability coefficient that is determined by assessing the degree of relationship between scores on two equivalent tests. She also recommends verifying the equivalence of test forms by means of establishing inter-rater reliability, i.e. the agreement between two or more independent raters or judges. The higher the percentage of agreement is, the more equivalent the tests are.

I used the two methods for verifying equivalence of both test forms (pre-test and post-test): expert evaluation by teachers and establishment of correlation between the test forms. The first version of tests was evaluated by two teachers from the Department of Foreign Languages at FEEC. Both gained a doctoral degree in English philology, and they have more than fifteen-year experience with ESP tests construction. Their task was to assess the equivalence of a form and content in each subtest using the evaluation tool where they specified the level of agreement or disagreement with the individual criteria of test forms on a 4-point Likert scale. The checklist included criteria for assessing equivalence of subtests, such as equivalent subtest format, task format, task instructions, number of items in each task, balanced representation of specialised vocabulary in each task, genre of reading text, genre of

recordings, length of text in the subtest Reading and length of recordings in the subtest Listening, micro-skills tested in the subtests Reading and Listening, time allocated for each subtest and an equivalent number of points for each task. To determine inter-rater reliability, I used the formula by Jackson (2009) where a number of agreements is divided by a number of possible agreements and then multiplied by 100.

Based on the first expert evaluation, I modified both test forms; in particular, I made the following changes: 1) clarifying the definitions in Task 1 of the subtest Use of English so that the answers were definite and/or including the other possible answers in the answer key; 2) reducing the length of text in the subtest Reading and 3) modifying the task instructions in Task 2 in the subtest Listening. Regarding the final forms of the test, the inter-rater reliability coefficient reached 0.99 for both the subtest Use of English and the subtest Listening, and 0.98 for the subtest Reading. From the expert evaluators' point of view, both test forms might be considered as equivalent.

I used the Spearman rank-order correlation coefficient r_s to establish correlation between both test forms. It measures the strength and direction of association between two ranked variables, in this case between students' pre- and post-test scores. The Spearman rank-order coefficient helps "quantify to what extent the two orders are similar, and therefore determine the strength of the association between two phenomena on which the orders were made" (Chráska, 2007, p. 103). The negative coefficient indicates the negative (opposite) association between ranks.

Chráska (2007) recommends using the Spearman rank-order correlation coefficient if a number of correlated paired ranks is not too high (max. about 30), and if less than four compared items (students) have the same rank. I included the test results of 23 students who took both test forms in the verification of test forms equivalence. This corresponds to Chráska's recommendation; however, more than four students had the same rank. The results in the form of each student's test scores and their ranks are presented in Appendix B.

After calculating the values from Appendix B, I got the Spearman rank-order correlation coefficient $r_s = \underline{0{,}85}$. The value of r_s indicates that there is a strong correlation between both test forms, which means that both test forms can be considered equivalent.

4.6 Determining Reliability of the Tests

To determine reliability of both the pre-test and the post-test, I used the split-half method, as it involves a single administration of a test, which seemed to be the best alternative because of the time constraints associated with this research. One of the other alternatives, test-retest method, when the same test is given to the same students within a short time interval, is impractical because "students may do better or worse the second time when they are accustomed to the test method, or are suffering from exhaustion or irritation" (Alderson et al., 1995, p. 87). If there were a longer time interval between test administrations, students would learn more of the language, which might also influence reliability.

The split-half method assesses the inter-item consistency of a test and measures the extent to which all parts of the test contribute equally to what is being measured. Inter-item consistency is a measure based on the correlations between different items on the same test (or the same subscale on a larger test) that determines whether several items that propose to measure the same general construct produce similar scores (Alderson et al., 1995; Field, 2013). This involves dividing a test into two, treating these two halves as being parallel versions, and correlating these two halves (Alderson et al., 1995). Chráska (1999, 2007) remarks that this method is suitable for testing a small number of students (no more than 30), but its disadvantage is that by reducing the number of test items reliability is reduced as well.

Moreover, the tests must have an even number of items. The tests that I constructed and verified contain a rather large number of items (72 in total), so its division into two parts should not influence reliability to a great extent. When splitting the test into halves, it is also necessary for the content of both halves to be equivalent. Bachman (2004) asserts that in some cases halves of a test measure different knowledge or skills. Accordingly, it is suggested that the test is divided into even and odd items while odd items will represent the first half of the test and even ones the second half. I constructed this test so that the same aspects could be measured in both halves since each subtest consists of the even number of items: the subtest Use of English contains 40 items, the subtest Reading has 18 items and the subtest Listening includes 14 items (see Appendix A). According to Chráska (2007), the reliability coefficient should be at least 0.8. A more detailed interpretation of the reliability coefficient is provided in Appendix C.

To calculate the reliability coefficient I used the Spearman-Brown formula. After substituting the values from Appendix D into the formula, I got the correlation coefficient between the

two halves of the pre-test $r_s = 0{,}937$. After substituting the values from Appendix E, I got the correlation coefficient between the two halves of post-test $r_s = 0{,}688$.

The Spearman-Brown reliability coefficient for the pre-test is $r_{sb} = 0{,}97$ and the Spearman-Brown reliability coefficient for the post-test is $r_{sb} = 0{,}82$. Regarding the fact that for educational assessment the reliability coefficient should be at least 0.8, both test forms (the pre-test and the post-test) are considered reliable.

The final version of the pre- and post-test includes: a task sheet, an answer sheet, and an answer key. I administered the pre-test at the beginning of the course and after the end of the course the students did the post-test. I provided the students with the results of the tests so that they could see their progress after completing the course English for IT.

5. Results

The aim of this section is to describe and compare students' pre- and post-test results and thus to answer the research questions. First, I evaluated the students' overall test scores by means of frequency distribution of individual scores. Appendix F shows frequencies n_i for both the pre-test and the post-test including calculated relative frequency f_i (expressed as a percentage). Appendix F also shows cumulative frequencies that equal the total of a frequency in the particular row and all frequencies in the previous rows. The table includes the values necessary for calculating the mean $\bar{x}$ and the standard deviation s.

I used the paired-samples t-test for a significance level of $p \leq 0.05$. This test is used when there are two experimental conditions (e.g. pre- and post-testing) and different participants were assigned to each condition (Field, 2013, Hendl, 2009). Each value $x_{pre\text{-}test}$ of the first sample has the corresponding value $x_{post\text{-}test}$ in the second sample. First, I calculated the degrees of freedom. In this case, the degrees of freedom were $f = 91$. The critical value for a significance level of $p \leq 0.05$ and 91 degrees of freedom was $t_{0,05}(91) = +/- 1{,}987$.

I verified all calculations in *IBM SPSS Statistics* (*Version* 25.0). The following section presents descriptive statistics of the discovered level of students' knowledge and skills in the pre- and post-test.

5.1 Description and Comparison of Students' Pre- and Post-Test Results

The relative frequency f_i in Appendix F provides information about the percentage of students who obtained the particular score in the pre-test and post-test. For example, the score of 62 points was obtained by 1.09 % of students in the pre-test and 7.61 % of students in the post-test. The cumulative frequency reveals that, for example, 10 students out of a total of 92 obtained 31 points at the maximum in the pre-test and 51 points at the maximum in the post-test.

Frequency distributions of pre- and post-test scores are illustrated in the histograms (blue for the pre-test and red for the post-test) in Figure 1. The histograms show that the higher scores were more frequent in students' post-tests than in their pre-tests.

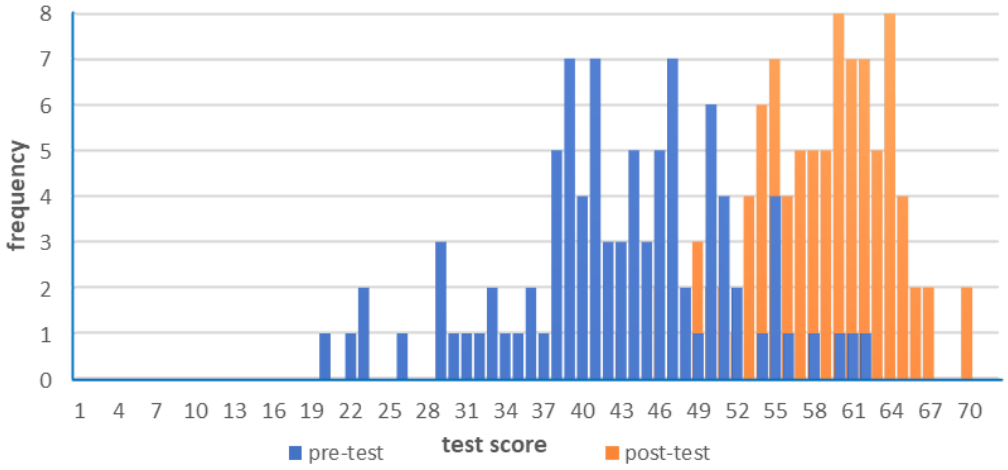

Figure 1. Frequency Histograms of Pre-Test (blue graph) and Post-Test (red graph) scores

After substituting the values from Appendix F, I got the means $\bar{x}_{pre\text{-}test}$ = 42.84, $\bar{x}_{post\text{-}test}$ = 58.53 and the standard deviations $s_{pre\text{-}test}$ = 8.71, $s_{post\text{-}test}$ = 5.42 for both the pre-test and the post-test. The students obtained the pre-test average score of 42.84 points, while in the post-test they achieved the average score of 58.53 points out of a total of 72 points.

Appendix G shows each student's pre- and post-test scores and test score difference x_d. As Appendix G indicates, all students obtained a higher score in the post-test than in the pre-test. The biggest difference (+31 points) is perceived at the student number 51 and the smallest difference (+2) is perceived at the student number 68. Based on the cut-off score (≥ 70 %, which is ≥ 50 points), the values indicate that 86 students succeeded and only 6 students failed in the post-test in comparison with the pre-test, where 22 students were successful and 70 failed (students who succeeded are highlighted in bold).

The sample mean of the differences was $\bar{x}_d$ = 15.70. After substituting the values from Appendix G, I got the sample standard deviation of the differences s_d = 0.68. The students' average post-test score (58.53) was higher than the average pre-test score (42.84). The *t*-test statistic value (t = 23.11) was greater than the critical value $t_{0,05}$ (91) = +/- 1.987; therefore, this difference can be considered as statistically significant. Similarly, the means, standard deviations and *t*-test statistic were calculated for the students' pre- and post-test scores in each of the subtests. The calculated *t*-test statistic for the subtest Use of English was 16.66 ($\bar{x}_{UoEpre\text{-}test}$ = 22.16, $\bar{x}_{UoEpost\text{-}test}$ = 30.99, $\bar{x}_{dUoE}$ = 8.83), which was higher than the critical value, so the difference is statistically significant. The *t*-test statistic for the subtest Reading was 13.52 ($x_{Rpre\text{-}test}$ = 12.13, $\bar{x}_{RPpost\text{-}test}$ = 15.29, $\bar{x}_{dR}$ = 3.16), which was also higher than the critical value. Regarding the last subtest Listening, the *t*-test statistic was 16.22 ($x_{Lpre\text{-}test}$ = 8.54, $\bar{x}_{LPpost\text{-}test}$ = 12.26, $\bar{x}_{dL}$ = 0.23), which was also higher than the critical value. Thus, this difference can be considered as statistically significant too.

Comparison of the students' pre- and post-test average score in the subtest Use of English, Reading and Listening is illustrated in the graph in Figure 2. The most considerable progress may be observed in the subtest Listening (the difference between the pre-test and the post-test was 26.57 %); in the subtest Use of English the difference was 22.08 % and the smallest progress (the difference 17.55 %) was made by students in the subtest Reading.

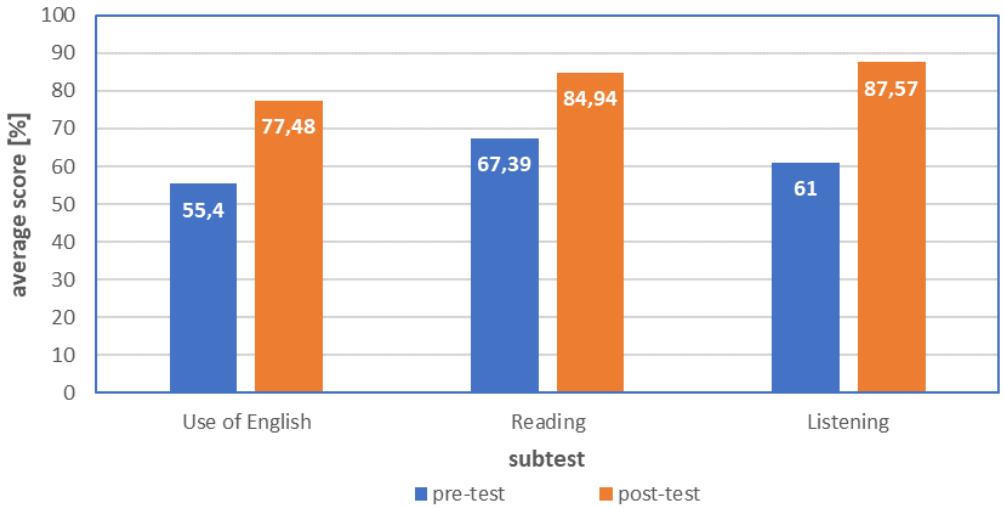

Figure 2. Comparison of Students' Average Pre- and Post-Test Scores in each Subtest (expressed as a percentage)

An important aspect in the evaluation of students' test results is also their success rate in the whole test and in its subtests. Table 2 shows the students' success rate in each subtest and in

the whole test.

Table 2: Success rate of students in different tested parts of the pre- and post-test.

Tested part	pre-test		post-test	
	success rate ≥ 70 %	% out of 92 students	success rate ≥ 70 %	% out of 92 students
Use of English	16	17.39	73	79.35
Reading	48	52.17	82	89.13
Listening	27	29.35	84	91.30
Whole test	22	23.91	86	93.48
Whole test (each subtest ≥ 70 %)	6	6.52	64	69.57

Based on the cut-off score ≥ 70 %, the values (highlighted in bold) in Table 2 indicate that the percentage of successful students in both the whole test and its each subtest (the condition was that the cut-off score in each subtest is ≥ 70 %) were only 6.52 % in the pre-test, while in the post-test the percentage was quite high – 69.57 %. If the determined success-rate (cut-off score ≥ 70 %) regarded the whole test only (without the requirement to succeed in each subtest), the percentage of successful students would be 93.48 % of students in the post-test and 23.91 % of students in the pre-test. In the post-test, the students were most successful (91.30 %) in the subtest Listening and least successful (79.35 %) in the subtest Use of English, where their success-rate was quite low even in the pre-test (only 17.39 %).

6. Discussion

In this section, I attempt to discuss the quality of the ESP coursebook pilot version and the changes that need to be made in its design based on students' pre- and post-test results and to formulate the preliminary design principles resulting from the first realization phase.

A comparison of students' pre- and post-test results in individual subtests and their success rates indicates which knowledge and skills were acquired by means of using the coursebook most and which were acquired to a limited extent. The percentage difference between the average pre- and post-test score was 21.79 %.

The greatest statistically significant difference between the pre-test and the post-test was noticed in the subtest Listening which, reached 26.57 %. A certain disagreement between the students' requirements for more difficult listening tasks resulting from the survey (see Ellederová, 2018) and their significant progress in listening skills can be observed here.

Students' success rate in this subtest might be ascribed to a relatively large number and variety of listening tasks in the coursebook (at least one task in each unit of the coursebook).

In the subtest Use of English, the difference between the students' pre-test and post-test scores was also statistically significant; the difference was quite big – 22.08 % . A range of specialized vocabulary highlighted in the text as well as a variety of vocabulary tasks often accompanied by the visuals in the coursebook probably helped students to succeed in this subtest. Highlighting the key words in the text encourages noticing which is "the essential starting point" for vocabulary acquisition (Lightbown & Spada, 2006, p. 115). An advantage of using visuals is that students can see an "instance of the meaning and this is likely to be remembered" (Nation, 2000, p. 126). Visuals as well as highlighted specialized vocabulary were also regarded as one of strengths according to the survey. Regarding the students' success rate, the main reason why students were least successful in the subtest Use of English of the post-test is probably because the students must acquire quite a wide professional vocabulary that is entirely new for them including linguistic means for expressing different language functions.

Even though the difference between the students' pre-test and post-test scores in the subtest Reading was statistically significant too, the difference between the average score was the smallest (17.6%). The reason for this relatively minor difference might be that the students of IT use reading skills quite often within their specific domain of study and work. IT professionals are a specific discourse community whose main goal is to transmit scientific information (for more details see Halliday, 2004), and their most frequent way of obtaining information is reading different scholarly texts (coursebooks, scientific books, research reports, hardware and software technical documentation and specification, hardware and software manuals) (Ellederová, 2020). Therefore, the students had probably developed their reading skills to a great extent even before they entered the course English for IT, which is supported by the fact that in the pre-test, the students were most successful in the subtest Reading (52.17%).

The research results indicate that some aspects of the coursebook should be modified by adding new texts and tasks or adapting the current ones. Based on the synthesis of the results from both stages of the research (pre- and post-testing and the survey), I will have to make the following modifications of the coursebook: 1) add more tasks for the acquisition of linguistic

means for expressing different language functions; 2) include more material for recycling and reinforcement focused in particular on vocabulary practice; 3) increase the level of difficulty of listening passages and add more complex tasks which will enable students to develop their listening skills and strategies (this last point especially to meet the students' requirements that the survey revealed).

Based on the results, I have developed a set of preliminary design principles. First, the main goal in creating an ESP coursebook should be to include such tasks that primarily ensure the integration of individual skills within one task, i.e. they should support the development of more than one skill (Hinkel, 2006; Mishan & Timmis, 2015) and at the same time contribute to the acquisition of new vocabulary or linguistic means. An example of such integration are the activities from the coursebook *English for Information Technology* where the listening tasks contribute not only to the development of listening skills but also to the development of speaking skills as well as consolidating new vocabulary through the following pair or group discussions about the issue.

Next, each unit in the ESP coursebook should be structured around the vocabulary definitions related to its topic. Vocabulary occurs systematically in professional lectures for non-native speakers, and therefore the "discourse role of definitions underlines the point that knowing the technical vocabulary is very closely related to knowing the subject area" (Nation, 2000, p. 323). The wordlist with phonetic transcription at the end of each unit in the coursebook allows students to determine which words and technical terminology they need to focus on and which words they may encounter in the final test.

Another important design principle is frequent repetition of keywords within a given unit and their occurrence in the following units. As Nation (2000) points out, "if a particular word occurs only once then it may be a burden but if it is repeated several times in the book then the initial learning effort is repaid by the opportunity to use that learning again when the word reoccurs" (p. 329). Vocabulary acquisition is also closely linked to the development of receptive and productive skills.

Tasks for reading, listening, and speaking in ESP coursebooks should be designed so that while doing them, students can use and gradually acquire vocabulary related to the particular topic. This is in accordance with the Involvement Load Hypothesis of Hulstijn and Laufer

(2001) who explain that vocabulary acquisition depends on the involvement load of a task, i.e. the amount of need, search, and evaluation it imposes. When presenting vocabulary in the ESP coursebook, it is also necessary to respect students' different learning styles through graphically highlighting key vocabulary in the text and presenting vocabulary using visuals and audio or video recording.

Lastly, the tasks and activities in the ESP coursebook should enable students to acquire linguistic means for expressing language functions within various roles and contexts in the field of information technology. In the coursebook, the sections Language Functions are based on the Present Practice Produce (PPP) approach. At the beginning of each section, the use of a given language function and a list of linguistic means is briefly explained, including examples of sentences in which the key structure is highlighted. This is be followed by a task in which students fill in or match various expressions to a text or image, or according to listening. Finally, students can acquire the linguistic means in the form of tasks and activities aimed at the development of speaking skills.

7. Conclusion

In this paper, I described the first two phases of DBR of the ESP coursebook *English for Information Technology*, and discussed the research findings and their impact on the coursebook redesign. Based on the results, I can conclude that the coursebook enabled students to improve their professional English knowledge as well as reading and listening skills. Students' pre-testing at the beginning of the course also proved to be beneficial (the positive impact of the washback effect) because students were acquainted with the level of language knowledge and skills expected from them in the final assessment of the course, which resulted in their success in post-testing.

Evaluation of the coursebook by means of students' pre-testing and post-testing has also certain limitations. The pre- and post-tests results could be influenced by the fact that during post-testing students were likely motivated by the pressure on their performance because the post-tests were part of their final assessment, while during pre-testing they were aware of the fact that their results would not influence their final grades. On the other hand, the post-tests may also have worsened students' results due to their anxiety during the final assessment. Another limitation may be the fact that I am the only teacher and supervisor of the course

English for IT where the coursebook is used. The ways in which the other teachers would use the coursebook in lessons might significantly influence students' performance in the post-tests. Even though the development of speaking skills provided by different tasks in the coursebook was evaluated by teachers and students in the survey (see Ellederová, 2018), students' speaking skills before and after using the coursebook were not tested due to time and organizational constraints. Assessing speaking skills of 92 students at the beginning and at the end of the course would be rather time-consuming. The number of lessons in the course English for IT would have to be shortened within the semester (it has already been shortened because of pre-testing) and at least four teachers would have to participate in pre-testing and post-testing speaking skills. Moreover, it would require changes in the timetable of the other courses. However, evaluation of the coursebook quality by means of testing students' speaking skills might be suggested for further research.

The preliminary design principles discussed in this study represent the fundamentals for the development of theories related to ESP coursebooks design. After the second realization phase of the coursebook DBR, I will see if and to what extent the modifications that resulted from the first realization phase were effective. Moreover, I will refine and advance ESP coursebooks design theories.

DBR requires iterative cycles of the stages, which will provide the opportunity to reflect and establish what dimensions of each intervention were "non-negotiable" or essential components at the core of each intervention that could not be changed. Therefore, the second realization stage will involve iteration, i.e. redesign of the coursebook, its repeated implementation (evaluation by teachers and students, pre-testing and post-testing of students) and the second data analysis and interpretation. Finally, the third realization stage will include production of the final design principles.

One of the main benefits of the ESP coursebook DBR presented in this study is that it should provide empirically grounded design principles and describe a research process that allows ESP teachers as authors of learning materials to learn instantly about the quality of their product and consequently incorporate their knowledge into practice.

References

Alderson, J. C. (1988). Testing and its administration in ESP. In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (pp. 87–97). Oxford: The British Council.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington D.C.: American Educational Research Association.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press

Bakker, A., & Van Eerde, H. A. A. (2013). An introduction to design-based research with an example from statistics education. In A. Bikner-Ahsbahs, C. Knipping, & N. Presmeg (Eds.), *Doing qualitative research: methodology and methods in mathematics education* (s. 429–466). New York: Springer.

Barab, S., & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences, 13*(1), 1–14.

Barnard, R., & Zemach, D. (2014). Materials for specific purposes. In B. Tomlinson (Ed.), *Developing materials for language teaching* (pp. 306–323). London: Bloomsbury.

Baumann, J., Blachowicz, C., Bates, A., Cieply, C., Manyak, P., Peterson, H., & Graves, M. (2013). The development of a comprehensive vocabulary instruction program for nine- to eleven-year-old children using a design experiment approach. In T. Plomp & N. Nieveen (Eds.), *Educational design research. Part B: Illustrative cases* (pp. 23–47). Enschede: SLO – Netherlands Institute for Curriculum Development.

Chráska, M. (1999). *Didaktické testy*. Brno: Paido.

Chráska, M. (2007). *Metody pedagogického výzkumu. Základy kvantitativního výzkumu.* Praha: Grada.

Council of Europe. (2011). *Common European framework of reference for languages: Learning, teaching, assessment.* Strasbourg: Council of Europe.

Danaye, T. M., & Haghigi, S. (2014). Evaluation of ESP textbooks: Evidence from ESP textbook of computer engineering major. *International Journal of Research Studies in Language Learning, 3*(2), 55–68.

Davies, A. (2001). The logic of testing languages for specific purposes. *Language Testing, 18*(2), 133-147.

Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5-8.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

Douglas, D. (2013). ESP and assessment. In B. Paltridge & S. Starfield (Eds.), *The handbook of English for specific purposes* (pp. 365–393). Chichester: John Wiley & Sons, Inc.

Dudley-Evans, T., & St John, M. J. (1998). *Developments in English for specific purposes.* Cambridge: Cambridge University Press.

Ellederová, E. (2016). *English for information technology.* Brno: VUT FEKT.

Ellederová, E. (2018). An ESP coursebook evaluation: Teachers' and students' viewpoints. In V. Janíková & S. Hanušová (Eds.), *Research in foreign language pedagogy* (pp. 149–165). Brno: Masarykova univerzita.

Ellederová, E. (2020). Konstrukční výzkum učebnice pro výuku angličtiny pro specifické účely: Hodnocení pilotní verze učebnice. *Pedagogika, 70*(1), 69–96.

Field, A. (2013). *Discovering statistics using IBM SPSS Statistics.* London: Sage.

Habtoor, H. A. (2012). English for specific purpose textbook in EFL milieu: An instructor's perspective evaluation. *International Journal of Linguistics, 4*(3), 44–59.

Halliday, M. A. K. (2004). *The language of science.* New York: Continuum.

Hendl, J. (2009). *Přehled statistických metod. Analýza a metaanalýza dat.* Praha: Portál.

Hinkel, E. (2006). Current perspectives on teaching the four skills. *TESOL Quarterly, 40*(1), 109–131.

Hulstijn, J. H., & Laufer, B. (2001). Some empirical evidence for the involvement load hypothesis in vocabulary acquisition. *Language Learning, 51*(3), 539–558.

Hymes, D. (1974). *Foundations of sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania.

Jackson, S. L. (2009). *Research methods and statistics. A critical thinking approach.* Belmont: Wadsworth Cengage Learning.

Jacoby, S., & McNamara, T. F. (1999). Locating competence. *English for Specific Purposes, 18*(3), 213–241.

Jafarabadi, M. N. S., Zarghi, N., Zolfaghari, V, & Kargozari, M. R. (2014). The effect of washback on reading comprehension of medical students in English for specific purposes classes. *Future of Medical Education Journal, 4*(4), 28–31.

Jebahi, K. (2009). Using a commercially developed ESP textbook: A classroom dilemma. *The Asian ESP Journal, 5*(2), 75–92.

Lightbown, P. M., & Spada, N. (2006). *How languages are learned.* Oxford: Oxford University Press.

Mishan, F., & Timmis, I. (2015). *Materials development for TESOL*. Edinburgh: Edinburgh University Press Ltd.

Mol, H., & Tin T. B. (2008). EAP materials in New Zealand and Australia. In B. Tomlinson (Ed.), *English language learning materials. A critical review* (pp. 74–92). London:

Continuum.

Nation, I. S. P. (2000). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Pearson, I. (1988). Tests as levers for change (or 'putting first things first). In D. Chamberlain & R. J. Baumgardner (Eds.), *ESP in the classroom: Practice and evaluation* (pp. 98–107). Oxford: The British Council.

Plomp, T., & Nieveen, N. (Eds.). (2013). *Educational design research*. *Part A: An Introduction.* Enschede: SLO – Netherlands Institute for Curriculum Development.

Razmjoo, S. A., & Raissi, R. (2010). Evaluation of SAMT ESP textbooks for the students of medical sciences. *Asian ESP Journal, 6*(2), 107–149.

Reeves, T. C., & Amiel, T. (2008). Design-based research and educational technology: Rethinking technology and the research agenda. *Educational Technology & Society, 11*(4), 29-40.

Rinder, J., Geslin, T. S., & Tual, D. (2016). A framework for language and communication in the CDIO syllabus. In J. Rinder (Ed.), *Proceedings of the 12th International CDIO Conference* (pp. 17–32). Turku: Turku University of Applied Sciences. Retrieved from http://www.cdio.org/files/document/cdio2016/72/72_Paper_PDF.pdf

Scio. (2014). *Teorie a metodika testů.* Retrieved from https://www.scio.cz/o-vzdelavani/teorie-a-metodika-testu/

Trna, J. (2011). Konstrukční výzkum v přírodovědných praktikách. *Scientia in Educatione, 2*(1), 3-14.

Van der Akker, J. (Ed.). (2006). *Educational design research.* Abingdon: Routledge.

Vičič, P. (2011). Preparing materials for ESP teaching. *Inter Alia 2*, 107–120.

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

Zangani, E. (2009). The ESP textbook problem: The evaluation of ESP textbooks in humanities in the undergraduate program of Iranian universities. *The Asian ESP Journal, 5*(2), 93–106.

Appendix A

Specification of the Test in ESP Focused on IT

Test type	*criterion-referenced test* including elements of a *proficiency test*		
Test objective	to test use of linguistic means (predominantly lexical items) and receptive skills (reading and listening) in ESP specialised in IT		
Language input characteristic	professional English language including specialised terminology from an IT sector; language level B2 defined as *Vantage*, *Limited Operational Proficiency*, *Upper-Intermediate* by *CEFR* and *GELS Framework*		
TLU domain characteristic	authentic tasks in the authentic environment (a discussion with an expert in programming languages, a discussion in a data centre, radio talk show with a software engineer, an interview with an administrator of websites specialised in gaming, a lecture) requiring reading comprehension of different genres (textbooks, scientific article, review in a scholarly journal) and listening comprehension (dialogue, discussion, lecture, presentation)		
Subtests	Subtest 1: Use of English	Subtest 2: Reading	Subtest 3: Listening
Time allocated	30 minutes	25 minutes	20 minutes
	Total time allocated is 80 minutes (5 minutes for instructions + 75 minutes test)		
Total number of points	40 points (cut-off score ≥ 28 points)	18 points (cut-off score ≥ 13 points)	14 points (cut-off score ≥ 10 points)
	Total number of points:72 ; cut-off score ≥ 50 points (≥70 %)		
Task types, number of items and their characteristic			
Use of English	Task 1	Task 2	Task 3
Task types and a number of items	short open answers; 15 items for nouns filling according to their definitions	short open answers; 3 items for verbs filling according to their definitions	matching the text with the correct word; 20 items for filling words in the text
Task focus	testing the specialised vocabulary acquisition		comprehension of both the text and the individual sentences structure, specialised vocabulary acquisition
Topics	introduction to IT and careers in IT (10 %); personal computer, types of computers, motherboard (20 %); input, output and storage devices (30%); software and Windows basics (10 %); networking, Internet access, World Wide Web and Internet safety (30 %)		databases, domain squatting, software security
Genre of a text in Task 3	textbook: Vermaat, M. E., et al. (2017). *Discovering Computers Enhanced: Tools, Apps, Devices, and the Impact of Technology.* Boston: Cengage Learning.		
Reading	Task 1	Task 2	Task 3
Task types and a number of items	6 items for the cloze text with drag-and-drop	6 items for matching	6 True/False items
Task focus	understanding the text and individual sentences structure (cohesion and coherence in professional discourse), testing acquisition of functional "reading" vocabulary	reading for specific information; testing specialised vocabulary acquisition	inferential reading comprehension and reading for specific information
Genre	scientific articles and reviews from *PC Magazine*		
Topics	data media, TCP/IP protocols, user interfaces		
Listening	Task 1	Task 2	
Task types and a number of items	7 True/False items, 3 items for gap filling	4 True/False items	
Genre	Interview	presentation/lecture/dialogue/discussion	
Task focus	ability to follow the main points, to infer links and connections and to detect specific information	ability to follow the main points and detect the specific information	

Topics	data storage and management, future of software technologies	RFID chips, computer games

Appendix B

Comparison of Students' Pre-Test and Post-Test Scores and Ranks

Student	Pre-test score	Rank in the pre-test	Post-test score	Rank in the post-test	Square of rank difference
A	62	1	65	1,5	0,25
B	61	2	64	4	4
C	58	3	64	4	1
D	56	4	60	8,5	20,25
E	55	5	64	4	1
F	50	6,5	59	10,5	16
G	50	6,5	60	8,5	4
H	49	8	61	7	1
I	47	9,5	65	1,5	64
J	47	9,5	55	12,5	9
K	46	11	55	12.5	2,25
L	43	12,5	59	10,5	4
M	43	12,5	53	15,5	9
N	41	14,5	53	15,5	1
O	41	14,5	54	14	0,25
P	40	16	49	20	16
Q	38	17	51	17	0
R	36	18	62	6	144
S	33	19	50	18	1
T	30	20	49	20	0
U	29	21	46	22	1
V	22	22	43	23	1
W	20	23	49	20	9
Σ	997		1290		309

Appendix C

Reliability Coefficient Interpretation Guideline

Reliability coefficient	Interpretation of a test
1,0 – 0,9	satisfactory for making decisions solely on its basis (e.g. regarding admission of students to the secondary school or university)
0,9 – 0,8	satisfactory as one of the documents for decision making
0,8 – 0,6	at the individual level unsatisfactory for decision making, but satisfactory for decision making regarding small groups (up to 10 people)

Note: Adapted from Scio (2014).

Appendix D

Students' Scores in Both Halves of the Pre-Test

Student	Total score	x_O	x_E	$x_O \bullet x_E$	x_O^2	x_E^2
A	62	32	30	960	1024	900
B	61	31	30	930	961	900
C	58	29	29	841	841	841
D	56	27	29	783	729	841
E	55	28	27	756	784	729
F	50	25	25	625	625	625
G	50	27	23	621	729	529
H	49	26	23	598	676	529
I	47	25	22	550	625	484
J	47	26	21	546	676	441
K	46	25	21	525	625	441
L	43	22	21	462	484	441
M	43	23	20	460	529	400

N	41	21	20	420	441	400
O	41	24	17	408	576	289
P	40	21	19	399	441	361
Q	38	21	17	357	441	289
R	36	19	17	323	361	289
S	33	18	15	270	324	225
T	30	16	14	224	256	196
U	29	17	12	204	289	144
V	22	16	6	96	256	36
W	20	12	8	96	144	64
Σ	997	531	466	11454	12837	10394

Appendix E

Students' Scores in Both Halves of the Post-Test

Student	Total score	x_O	x_E	$x_O \bullet x_E$	x_O^2	x_E^2
A	65	33	32	1056	1089	1024
B	64	32	32	1024	1024	1024
C	64	33	31	1023	1089	961
D	60	29	31	899	841	961
E	64	30	34	1020	900	1156
F	59	27	32	864	729	1024
G	60	32	28	896	1024	784
H	61	30	31	930	900	961
I	65	34	31	1054	1156	961
J	55	29	26	754	841	676
K	55	28	27	756	784	729
L	59	33	26	858	1089	676
M	53	26	27	702	676	729
N	53	28	25	700	784	625
O	54	26	28	728	676	784
P	49	25	24	600	625	576
Q	51	27	24	648	729	576
R	62	32	30	960	1024	900
S	50	26	24	624	676	576
T	49	26	23	598	676	529
U	46	25	21	525	625	441
V	43	21	22	462	441	484
W	49	27	22	594	729	484
Σ	1290	659	631	18275	19127	17641

Appendix F

Students' Scores and Their Frequency in the Pre-Test and the Post-Test

Score x_i	Frequency n_i		Relative frequency f_i		Cumulative frequency		$n_i \bullet x_i$ *pre-test*	$n_i \bullet x_i$ *post-test*	$n_i (x_i - \bar{x})^2$ pre-test	$n_i (x_i - \bar{x})^2$ post-test
	pre-test	post-test	pre-test	post-test	pre-test	post-test				
0	0	0	0.00	0.00	0	0	0	0	0.00	0.00
1	0	0	0.00	0.00	0	0	0	0	0.00	0.00
2	0	0	0.00	0.00	0	0	0	0	0.00	0.00
3	0	0	0.00	0.00	0	0	0	0	0.00	0.00
4	0	0	0.00	0.00	0	0	0	0	0.00	0.00
5	0	0	0.00	0.00	0	0	0	0	0.00	0.00
6	0	0	0.00	0.00	0	0	0	0	0.00	0.00
7	0	0	0.00	0.00	0	0	0	0	0.00	0.00
8	0	0	0.00	0.00	0	0	0	0	0.00	0.00
9	0	0	0.00	0.00	0	0	0	0	0.00	0.00
10	0	0	0.00	0.00	0	0	0	0	0.00	0.00
11	0	0	0.00	0.00	0	0	0	0	0.00	0.00
12	0	0	0.00	0.00	0	0	0	0	0.00	0.00

13	0	0	0.00	0.00	0	0	0	0	0.00	0.00
14	0	0	0.00	0.00	0	0	0	0	0.00	0.00
15	0	0	0.00	0.00	0	0	0	0	0.00	0.00
16	0	0	0.00	0.00	0	0	0	0	0.00	0.00
17	0	0	0.00	0.00	0	0	0	0	0.00	0.00
18	0	0	0.00	0.00	0	0	0	0	0.00	0.00
19	0	0	0.00	0.00	0	0	0	0	0.00	0.00
20	1	0	1.09	0.00	1	0	20	0	521.53	0.00
21	0	0	0.00	0.00	1	0	0	0	0.00	0.00
22	1	0	1.09	0.00	2	0	22	0	434.18	0.00
23	2	0	2.17	0.00	4	0	46	0	787.01	0.00
24	0	0	0.00	0.00	4	0	0	0	0.00	0.00
25	0	0	0.00	0.00	4	0	0	0	0.00	0.00
26	1	0	1.09	0.00	5	0	26	0	283.48	0.00
27	0	0	0.00	0.00	5	0	0	0	0.00	0.00
28	0	0	0.00	0.00	5	0	0	0	0.00	0.00
29	3	0	3.26	0.00	8	0	87	0	574.39	0.00
30	1	0	1.09	0.00	9	0	30	0	164.79	0.00
31	1	0	1.09	0.00	10	0	31	0	140.11	0.00
32	1	0	1.09	0.00	11	0	32	0	117.44	0.00
33	2	0	2.17	0.00	13	0	66	0	193.53	0.00
34	1	0	1.09	0.00	14	0	34	0	78.09	0.00
35	1	0	1.09	0.00	15	0	35	0	61.42	0.00
36	2	0	2.17	0.00	17	0	72	0	93.49	0.00
37	1	0	1.09	0.00	18	0	37	0	34.07	0.00
38	5	0	5.43	0.00	23	0	190	0	116.98	0.00
39	7	0	7.61	0.00	30	0	273	0	103.06	0.00
40	4	0	4.35	0.00	34	0	160	0	32.19	0.00
41	7	0	7.61	0.00	41	0	287	0	23.62	0.00
42	3	0	3.26	0.00	44	0	126	0	2.10	0.00
43	3	1	3.26	1.09	47	1	129	43	0.08	241.27
44	5	0	5.43	0.00	52	1	220	0	6.76	0.00
45	3	0	3.26	0.00	55	1	135	0	14.04	0.00
46	5	2	5.43	2.17	60	3	230	92	50.02	314.15
47	7	0	7.61	0.00	67	3	329	0	121.31	0.00
48	2	0	2.17	0.00	69	3	96	0	53.31	0.00
49	1	3	1.09	3.26	70	6	49	147	37.98	272.63
50	6	2	6.52	2.17	76	8	300	100	307.85	145.62
51	4	2	4.35	2.17	80	10	204	102	266.54	113.49
52	2	1	2.17	1.09	82	11	104	52	167.92	42.68
53	0	4	0.00	4.35	82	15	0	212	0.00	122.46
54	1	6	1.09	6.52	83	21	54	324	124.61	123.29
55	4	7	4.35	7.61	87	28	220	385	591.75	87.37
56	1	4	1.09	4.35	88	32	56	224	173.26	25.66
57	0	5	0.00	5.43	88	37	0	285	0.00	11.75
58	1	5	1.09	5.43	89	42	58	290	229.92	1.42
59	0	5	0.00	5.43	89	47	0	295	0.00	1.09
60	1	8	1.09	8.70	90	55	60	480	294.57	17.22
61	1	7	1.09	7.61	91	62	61	427	329.89	42.60
62	1	7	1.09	7.61	92	69	62	434	367.22	84.14
63	0	5	0.00	5.43	92	74	0	315	0.00	99.77
64	0	8	0.00	8.70	92	82	0	512	0.00	239.10
65	0	4	0.00	4.35	92	86	0	260	0.00	167.29
66	0	2	0.00	2.17	92	88	0	132	0.00	111.51
67	0	2	0.00	2.17	92	90	0	134	0.00	143.38
68	0	0	0.00	0.00	92	90	0	0	0.00	0.00
69	0	0	0.00	0.00	92	90	0	0	0.00	0.00
70	0	2	0.00	2.17	92	92	0	140	0.00	262.98
71	0	0	0.00	0.00	92	92	0	0	0.00	0.00
72	0	0	0.00	0.00	92	92	0	0	0.00	0.00
Σ	92	92	100	100			3941	5385	6898.51	2670.87

Appendix G

Comparison of Each Student's Pre- and Post-Test Scores

Student's number	$x_{pre\text{-}test}$	$x_{post\text{-}test}$	x_d	$(x_d - \bar{x})2$
1	**62**	**65**	+3	161.18

2	**61**	**64**	+3	161.18
3	**58**	**64**	+6	94.01
4	**56**	**60**	+4	136.79
5	**55**	**64**	+9	44.83
6	**50**	**59**	+9	44.83
7	**50**	**60**	+10	32.44
8	49	**61**	+12	13.66
9	47	**65**	+18	5.31
10	47	**55**	+8	59.22
11	46	**55**	+9	44.83
12	43	**59**	+16	0.09
13	43	**53**	+10	32.44
14	41	**53**	+12	13.66
15	41	**54**	+13	7.27
16	40	49	+9	44.83
17	38	**51**	+13	7.27
18	36	**62**	+26	106.18
19	33	**50**	+17	1.70
20	30	49	+19	10.92
21	29	46	+17	1.70
22	22	43	+21	28.14
23	20	49	+29	177.01
24	**60**	**70**	+10	32.44
25	**55**	**67**	+12	13.66
26	38	**63**	+25	86.57
27	**50**	**70**	+20	18.53
28	**55**	**66**	+11	22.05
29	46	**67**	+21	28.14
30	39	**62**	+23	53.35
31	**52**	**62**	+10	32.44
32	39	**60**	+21	28.14
33	48	**63**	+15	0.48
34	39	**58**	+19	10.92
35	47	**64**	+17	1.70
36	39	**62**	+23	53.35
37	44	**64**	+20	18.53
38	**50**	**58**	+8	59.22
39	**51**	**62**	+11	22.05
40	44	**56**	+12	13.66
41	**55**	**64**	+9	44.83
42	47	**61**	+14	2.88
43	39	**63**	+24	68.96
44	39	**58**	+19	10.92
45	48	**60**	+12	13.66
46	36	**60**	+24	68.96
47	44	**57**	+13	7.27
48	47	**63**	+16	0.09
49	**51**	**66**	+15	0.48
50	41	**61**	+20	18.53
51	31	**62**	+31	234.22
52	34	**64**	+30	204.61
53	**50**	**61**	+11	22.05
54	41	**55**	+14	2.88
55	46	**61**	+15	0.48
56	**50**	**65**	+15	0.48
57	**52**	**64**	+12	13.66
58	38	**55**	+17	1.70
59	38	**56**	+18	5.31
60	44	**63**	+19	10.92
61	46	**61**	+15	0.48
62	47	**65**	+18	5.31
63	45	**57**	+12	13.66
64	42	**60**	+18	5.31
65	42	**57**	+15	0.48
66	41	**62**	+21	28.14
67	38	**59**	+21	28.14
68	**54**	**56**	+2	187.57
69	**51**	**61**	+10	32.44
70	44	**59**	+15	0.48

71	35	**60**	+25	86.57
72	**51**	**59**	+8	59.22
73	32	**58**	+26	106.18
74	47	**60**	+13	7.27
75	40	**56**	+16	0.09
76	40	**54**	+14	2.88
77	37	**53**	+16	0.09
78	29	**53**	+24	68.96
79	41	**54**	+13	7.27
80	43	**55**	+12	13.66
81	45	**54**	+9	44.83
82	26	**54**	+28	151.4
83	33	**55**	+22	39.74
84	46	**57**	+11	22.05
85	23	**52**	+29	177.01
86	40	**51**	+11	22.05
87	39	**50**	+11	22.05
88	42	**57**	+15	0.48
89	23	46	+23	53.35
90	41	**54**	+13	7.27
91	29	**58**	+29	177.01
92	45	**55**	+10	32.44
Σ			1444	3861.48