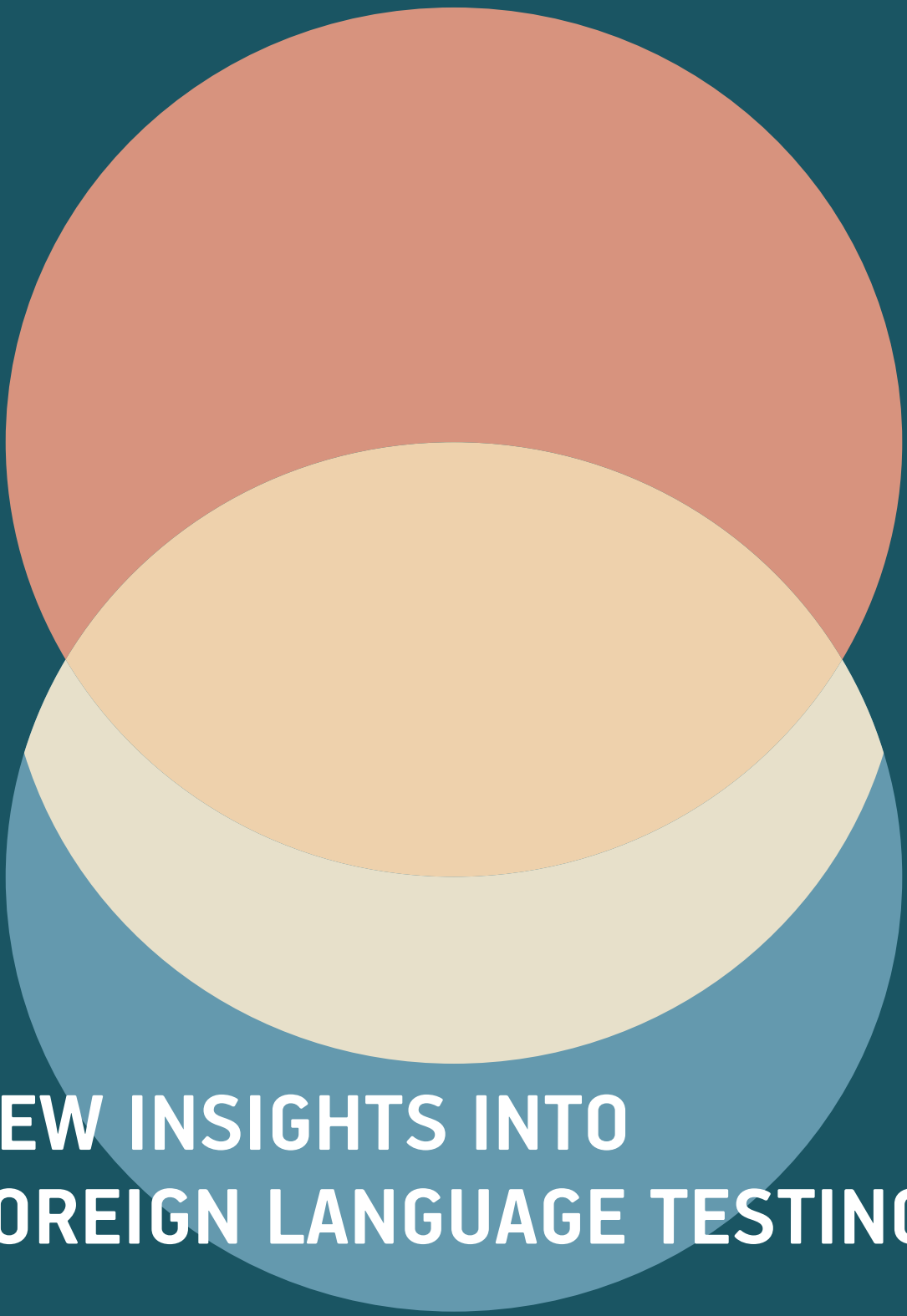
FLV | FAKULTET ZA PRAVNE I POSLOVNE STUDIJE DR LAZAR VRKATIĆ



NEW INSIGHTS INTO FOREIGN LANGUAGE TESTING

Tatjana Glušac

TATJANA GLUŠAC

NEW INSIGHTS INTO FOREIGN LANGUAGE TESTING

Novi Sad, 2022

To my three angels, Dušan, Vidak, and Nikola, whose thirst for knowledge drives my life mission to strive and fight for quality education.

CONTENTS

PREFACE

As the title suggests, the aim of this book is to offer insight into certain issues of interest in the domain of foreign language testing that have been undeservedly overlooked. In truth, the area of testing, including foreign language testing, is so vast that it is virtually impossible to cover all relevant issues in sufficient detail in a relatively short span of time or at one go, whether through a course, a publication, a seminar, or other means. Now, however, the time has come to discuss and explore more thoroughly certain issues in this context that have not been sufficiently addressed, especially in light of their prevailing importance today. These key issues include: the relevance of teacher-made tests as opposed to ready-made, commercial, or publishers' tests; the extent to which different levels of cognitive processing are present in foreign language tests; the question of whether foreign languages allow for the improvement of students' subject-specific critical thinking; and the effectiveness and quality of test task instructions.

This book primarily addresses teacher-made, or classroom, tests, i.e., those tests created by foreign language teachers themselves for their own measurement purposes. In this publication, a parallel is frequently drawn between these tests and the contrasting high-stakes tests to illustrate the differences in, and/or the distinct scopes of, the application of a certain testing principle to either type of test, thus enabling the reader to gain a greater understanding of the function of these tests and their corresponding operations.

Not only is it my intention to ground the issues discussed in this book in relevant theory, but also to illustrate the points tackled with appropriate

examples. In doing so, I hope I will clarify my points and explain how a certain issue discussed should or should not be used in practice.

First and foremost, this book is intended for the master's students in the Department of English, at the Faculty of Law and Business Studies *Dr Lazar Vrkatić*, Novi Sad, as part of the reading list these students are required to consult for the purposes of the course "Assessment in Foreign Language Teaching and Learning." It is also anticipated that this book will be of interest to and relevant for numerous other readers. For instance, practicing language teachers could use it as a reference to reinforce and expand their existing theoretical and practical knowledge on testing and enrich it with new insights. Numerous examples throughout the book are intended to help practicing teachers apply the suggestions and ideas presented in this book. Additionally, the book could be particularly useful to educators providing instruction and training to student-teachers, as it addresses issues regarded as highly important by modern literature, but which have yet to be covered in sufficient detail. Last but not least, the book could serve as a starting point for researchers aiming to further explore the topics presented here, since almost all the chapters in this publication are based on research I've personally carried out with colleagues.

This book is to a great extent a result of my many years of teaching foreign language assessment to English language master's students and leading seminars for practicing English language teachers in assessment and critical thinking in foreign language education. As well, the book is a result of my own inquiries, queries, and experiments with issues not sufficiently covered in relevant theoretical or empirical works. I am grateful to all my inquisitive students and to the many practicing teachers who have attended my seminars for raising various pertinent questions and sharing their diverse experiences during our meetings.

This book would not have come to be if it had not been for a number of people. First and foremost, I am indebted to Ivana Vrkatić, who granted me permission to publish this book. My heartfelt thanks also goes to my colleagues and coauthors of the papers that serve as the basis for some chapters of this book for giving me their consent to use the papers for the purposes of this book: Mira Milić, Vesna Pilipović, Isidora Wattles, and Nataša Bogdanov. Also, I am grateful to Ana Sentov, head of the Department of English at the Faculty of Law and Business Studies *Dr Lazar Vrkatić*, and all my department colleagues for supporting me in my writing. Special thanks also goes to a number of people who have provided me with valuable suggestions and insights both before and during the book's writing: Tvrtko Prćić, Faculty of Philosophy, University of Novi Sad; Mira Milić, Faculty of Sport and Physical Education, University of Novi Sad; Vesna Pilipović, Faculty of Law and Business Studies *Dr. Lazar Vrkatić*, Union University; Radmila Bodrič, Faculty of Philosophy, University of Novi Sad; and Çiler Hatipoğlu, Faculty of Education, Middle East Technical University, Turkey. I also want to thank Andrew Wiesike, Ferenc Finčur, and Aleksandar Međedović for ensuring the language and technical correctness of the publication. A special 'thank you' goes to my two very close friends, Vera Džodan and Aleksandra Relić, whose encouragement and words of wisdom meant a world to me at certain moments. Last but not least, I am grateful to my family for their continuous support and understanding.

Tatjana Glušac,
Novi Sad, September 2021

TERMONOLOGY AND TYPOGRAPHY USED

‘Assessment,’ ‘testing,’ and ‘evaluation’ are not used as synonyms. Their understanding and use in this book are defined in Chapter 2.

‘Commercial test,’ ‘ready-made test,’ and ‘(coursebook) publisher’s test’ are used as synonyms.

‘High-stakes test,’ ‘standardized test,’ and ‘large-scale test’ are used as synonyms.

‘Teacher-made test,’ ‘classroom test,’ and ‘low-stakes test’ are used as synonyms.

Italics are used for the terminology defined, emphasis, and parts of tasks analyzed in the text.

Bold is used for test task instructions, titles of chapters, and their sections.

The mark [Example] and a corresponding number is used to denote an example, where the no. is each time replaced with consecutive numbers.

The symbol […] is used in test task examples to indicate that only a part of the original task is given as an illustration, not the entire task.

1. THE CONCEPT OF TESTING

Testing is an integral part of teaching, and, what is more, testing contributes to the improvement of the quality of teaching and learning. It goes without saying that in order to enhance the quality of teaching and learning, educators need quality tests as precise and reliable measuring instruments. It is universally asserted in the field of foreign language that testing is a rather complex activity requiring numerous skills and substantial knowledge on the part of the tester. Moreover, particular authors (e.g., Bachman, 1995; Dimitrijević, 1999) have highlighted that, despite testing falling within the scope of different disciplines, such as applied linguistics, psychology, pedagogy, etc., there are still some unresolved questions, which makes the practice of test design challenging.

A test is a measuring instrument by means of which assessment and evaluation can be carried out for different purposes. Many authors (e.g., Bachman & Palmer, 2004; Brown, 2000; Heaton, 1990; Hughes, 2003; Martinez et al., 2009) have regarded *assessment* and *evaluation* to be synonymous terms, while others have made a distinction between the two (e.g., Abbas, 1994; Angelo & Cross, 1993; Hamp-Lyons, 2016; Hattie & Brown, 2010; Starr, 2014). The latter group of authors have expressed the belief that assessment is primarily done for formative purposes, i.e., while learning is still in progress, with the aim of helping learners master the content taught. In other words, assessment is done for the purpose of helping students 'form' the knowledge base of a particular subject or topic and it is conducted at different phases of learning. Also, assessment can be conducted with the aim of investigating the effects of applied teaching methods and techniques on the quality of

knowledge students are acquiring. Assessment can be done both explicitly and implicitly, while the result of an assessment does not need to be a grade or a number of points. For instance, if a teacher wonders how well the students have understood the last unit covered, he/she could carry out an informal quiz at the beginning of the class immediately following the class in which the new content was covered in order to decide whether he/she should reteach the unit, or elements of it, or whether he/she should proceed to the next stage of teaching. Such assessment is implicit, as students do not know they are being assessed, and the result of such a process is not a grade, but insight or feedback. Decisions made on the basis of information obtained through assessment do not necessarily impact students individually, but they can result in actions that impact the entire class, such as a changed teaching routine, the employment of different teaching techniques, re-teaching of the previously covered content, etc. Individual impact is possible, however, in situations such as pairing or grouping students.

When assessment is done for the purpose of judging the amount or quality of acquired knowledge at the end of a certain period of learning, the process conducted in such a context is *evaluation*. Evaluation, here, is almost exclusively a formal, explicit procedure, the aim of which is making a judgment that is then expressed in the form of a grade or a number of points. The information gathered through evaluation helps the teacher or other education authority make certain pedagogical decisions that are, more often than not, highly impactful for the relevant students. For instance, an entrance examination is evaluative in that a decision is made based on the entrance test results whether a student will be admitted to a university or not. Such a decision bears a great deal of significance for the student, as it directly impacts the trajectory of the student's life in one way or the other. In addition to its being performed constantly, regarding the order of the learning process, assessment should precede evaluation, as it is aimed at improving the process of learning before its final outcome is measured. As mentioned before, a

test is an instrument used both in assessment and evaluation, but both these processes may also be conducted by means of other measurement instruments, including quizzes, games, discussions, interviews, and portfolios, to name a few. Of all the procedures used for conducting both of these processes, testing is probably the most commonly employed, despite it indisputably requiring a great deal of relevant mastery and knowledge to properly construct and administer tests and interpret their results.

When assessment is conducted by means of a test, either the teacher constructs one himself/herself or he/she uses a ready-made test. Evaluation, in contrast, is typically conducted by means of a standardized or high-stakes test constructed by a group of individuals specialized in test design. A teacher-made test can also be used for the purpose of evaluation in situations in which the testing is done at the end of a learning period to measure the quality or amount of knowledge gathered over a course of time, such as a term or school year. In such situations, a grade is assigned as a form of final judgement regarding whether students have met the set learning goals or outcomes. On many occasions, the grade assigned in such a test is final and cannot be corrected, making the described process evaluation rather than assessment. Dimitrijević (1999, p. 61) adds to the discussion on terminology by saying that avoidance of the term *grading* is evident in such contexts since it bears negative connotations for students; instead, *evaluation* is used in order to diminish the students' fear of being graded.

Along the same lines, assessment is measurement *for* learning, while evaluation is measurement *of* learning. In other words, assessment is conducted while learning is still in progress in order to improve it (measurement for learning), while evaluation measures the product of learning (measurement of learning). As noted by Hattie and Brown (2010), the results obtained through evaluation can also be used for assessment purposes, or, more precisely, what one determines as the final product of students' learning can be integrated in reconceptualizing the respective course or its content when it is implemented the next time.

In this book, the two terms — *assessment* and *evaluation* — will not be used as synonyms, but as denoting processes that have distinctive and sometimes overlapping features. As defined by Hattie and Brown (2010):

> Assessment relates to the identification of characteristics of a trait, and evaluation relates to the establishment of value and worth of a product, process, person, policy or program. Assessment refers to 'What's so?' and evaluation to 'So what?' Both depend on high-quality measurement, and both focus on the qualities, degrees, and characteristics of student learning of the material deemed important by society and identified in the curriculum. The validity and reliability of such assessments and evaluations depend on our ability to specify what is to be learned and defensible measures of progress in each curriculum domain. (p. 103)

Language is a form of behavior and measuring any form of behavior is a rather complex undertaking. Knowledge of a foreign language entails a number of developed competences (linguistic, pragmatic, etc.), skills (speaking, reading, writing, listening), and a number of different types of knowledge (grammar, vocabulary, etc.). Therefore, in order to assess learners' knowledge of a foreign language, all the aforementioned components need to be both covered in class and assessed to determine the scope of their mastery.

Testing falls within the scope of several disciplines (applied linguistics, psychology, pedagogy, etc.) and to be competent in test design, one needs to be familiar with the current relevant and preeminent doctrines from all these fields, and not merely with those from the field(s) the test designer has specialized in. For this reason, a number of issues related to test design are likely to prove problematic for test designers whose perspective is relatively limited to a certain field. For instance, psychologists would probably be more successful in ensuring the metric characteristics of a test than foreign language teachers, while the latter group would probably be more successful in dealing with language-related aspects, such as the content of the test. Even though testing is explored within different disciplines, there are still some unresolved questions, which makes foreign language test design a challenge.

As Dimitrijević (1999) asserts, we still do not have a definite answer to the questions of how to choose a testing sample, whether a test measuring language competence can tell us what our students' language performance is, to what degree the results of a test administered in a classroom can reveal students' language performance in real-life situations, etc. Despite the apparent presence of unresolved questions whose answers and solutions researchers continue to seek, testing remains an integral part of quality teaching and learning and needs to be conducted. Even though many authors (e.g., Glušac & Milić, 2021; McMillan, 2000; Piggie & Marso, 1988) do acknowledge and confirm that tests made by teachers themselves often contain a number of flaws, they remain indispensable instruments for obtaining a realistic picture of what impact teaching and learning strategies and practices have on students. However, test designers need to understand that continuous professional development in test development and design is paramount in ensuring the creation of quality of tests, and so is continuous improvement/revision of relevant academic courses offered to future foreign language teachers.

Despite Hattie and Brown's (2010) assertion that a teacher's familiarity with students and the learning context enables him/her to conduct measurement in ways other than applying formal measurement instruments, i.e., tests, tests seem to be the most commonly used measuring instrument in the classroom (Frey & Schmitt, 2007; Glušac & Milić, 2021). However, the mere thought of a testing situation is more likely than not to exert affective block in students (Dimitrijević, 1999) for several reasons: teachers tend to place inordinate importance on tests and test scores; students may not be properly prepared for taking a particular test; a test may serve as means for punishing students; a test may not accurately reveal students' real knowledge; students may not be scored properly if the scoring system is flawed; the test may contain mistakes or ambiguities that prevent students from appropriately answering questions, etc. Marso and Piggie (1993) add to this the following: "[Un]announced tests, carelessly administrated tests, poorly monitored tests, and tests perceived by

pupils to be unfair not only adversely impact upon student performance but tend to heighten test anxiety and encourage cheating" (p. 133). Hattie and Brown (2010) acknowledge that errors are typical of classroom tests and state that "[g]iven the persistent contact teachers have with learners and the multiple opportunities that provide for of-the-moment assessment, it must be assumed that any error, mis-specification or randomness in the decisions and actions taken by the teacher should have little or no negative effect on the learner" (104). However, the situation is often quite the opposite. Teachers often lack training in test development and design and thus overlook the downsides and/or flaws of a test. They attach unreasonably high importance to test results (Marso & Piggie, 1993, p. 151) and rely on them when making both minor and major pedagogical decisions. Relatedly, certain authors (e.g., Hattie & Brown, 2010) have asserted that teachers require training in test results interpretation. In the same vein, teachers frequently possess a number of misconceptions in relation to testing, assessment, and evaluation (Bachman & Palmer, 2004) and some of them might stem from their improper understanding of the terms commonly associated with the measurement process (*testing*, *assessment*, *evaluation*), its aims, and its manner(s) of realization. The lack of consensus in relation to key terms also affects research in the field of classroom assessment (Frey & Schmitt, 2007, p. 402). Many authors around the world (Frey & Schmitt, 2007; Hattie & Brown, 2010; Hidri, 2021; Glušac & Milić, 2021; Tsagari et al., 2018) have called for the need to improve foreign language teachers' assessment literacy. An assessment literate teacher has adequate knowledge and skills needed for conducting effective measurement and using corresponding results to make informed decisions pertaining to teaching and learning.

Teachers' lack of assessment skills is a world-wide concern. Several years ago the need for enhanced assessment literacy of foreign language teachers was particularly recognized by a number of researchers from various European countries, including Norway, Greece, Hungary, Cyprus, Germany, and the UK. This belief united them in developing a three-year Erasmus+ Program project

financed by the European Commission whose aim was to create an infrastructure for helping teachers hone their assessment literacy skills. One of the results of the project is a website (www.taleproject.eu) that offers its visitors, intended to be foreign language teachers, a downloadable handbook, as well as an opportunity to take an eight-module self-access course in test design. This certainly is a unique opportunity and an invaluable source of theoretical and practical information.

In Serbia, foreign language assessment is taught to students studying English at four out of five university-level departments of English. Three departments offer the course at the undergraduate level, two as an obligatory course and one as an elective, while one department offers the course to master-level students as an elective. For the gaining of relevant knowledge and skills pertinent to this aspect of foreign language teaching, English language teachers, as well as teachers of other foreign languages, have few other possibilities to hone their assessment skills. Assessment, testing, and evaluation are topics covered in a number of accredited seminars for English language teachers in Serbia, but they do not appear to be given due attention; rather, they are listed among other numerous topics dealt with in wide-ranging seminars, for which reason they are almost certainly not covered in adequate detail. As research (e.g., Marso & Piggie, 1993) evidences, teachers acknowledge they need additional training to improve their test design skills, but they are unwilling to pursue it, most probably because what they require is practical rather than theoretical knowledge, which is what they typically get (Stiggins, 1988, cited in Marso & Piggie, 1993, p. 153). To this, Hidri (2021, p. 8) adds that quality programs which enhance assessment literacy skills need to be well-rounded and include training in both measuring and improving conceptions of assessment. Similarly, Hatipoğlu (2015) claims that "foreign language teacher training programs should monitor, revise and regularly innovate their English Language Testing and Evaluation (ELTE) courses so that they prepare future teachers better for the challenges of language assessment in their specific contexts" (p. 111).

The aim of this book is not to provide a review of the fundamental assessment-related issues that the publications of renowned authors in this field have presented. Due to their comprehensiveness and quality, as well as their authors' vast experience and expertise, these publications should comprise the list of required readings for all those who conduct, or will go on to conduct, any form of measurement in education. Rather, this book is a humble attempt to shed light on certain pertinent issues in foreign language testing that have thus far received insufficient scholarly attention. Firstly, by drawing on relevant literature, this book offers an explanation for why teacher-made tests, despite all their possible flaws, are still better measuring instruments than ready-made tests. Secondly, as testing is not regarded solely as a means of assessing knowledge, but also as a learning tool, this book offers research findings related to the degree to which test tasks invite and stimulate students to use their foreign language knowledge constructively and freely; or, in other words, whether such tasks require students to productively manipulate the knowledge they possess or to simply regurgitate it. Finally, the book offers a comprehensive literature review on test task instructions, a topic not covered systematically or in satisfactory detail in contemporary literature on test design.

1.1. Basic postulates of foreign language testing

Testing in general and foreign language testing in particular need to be based on a number of postulates if they are to be considered quality, reliable, and effective.

As put forward by Bachman and Palmer (2004), *there is no such a thing as 'the best' test,* or a test that could serve as a model for designing all subsequent tests. Every test needs to be created for a specific group of learners who have received instruction in a particular context and with a particular purpose in mind. This means that individual learners' characteristics need to be taken

into consideration when designing a test, such as their age, the context of learning, their preferences, backgrounds, etc. Proper consideration of these factors enables and potentially ensures the choosing of age-appropriate and culture-appropriate test techniques, topics of interest, and/or the variant(s) of a foreign language students might have been exposed to through instruction (e.g., whether they were taught British or American English or received instruction in English for specific purposes, etc.), as well as the avoidance of culturally sensitive topics. Moreover, each test should be created for a different purpose, the needs of which should be clear to the test designer during its inception. Based on the purpose and the construct, the test designer should decide what the test results will be used for and choose corresponding test techniques. The purpose of a test may be to assess students' ability to communicate for academic or business purposes, for instance, which should impact the type of test techniques in the sense that only those tasks revealing test takers' ability to communicate in the context(s) relevant to the designated purpose, such as writing for business or academic purposes, giving presentations, etc., should be included in the test. Other language components may be less relevant in such a test, or may be assessed indirectly, i.e., through writing or speaking. As Bachman and Palmer (2004) put it, "In order for a particular language test to be useful for its intended purposes, test performance must correspond in demonstrable ways to language use in non-test situations" (p. 9).

Bachman and Palmer (2004) and Dimitrijević (1999) claim that language *testing should relate both to language teaching and language use.* In other words, the way in which we test and the content tested need to correspond as closely as possible to the language instruction students have received prior to the test. A language teacher's beliefs regarding the language itself and the way it needs to be taught strongly impact his/her teaching practice. Namely, the answers to the questions such as How important is grammar for one's foreign language competence?, Does writing really contribute to learners' foreign language proficiency?, and the like impact the way a teacher teaches. Owing

to his/her own answers to such questions, the teacher chooses the content to teach, opts for certain teaching techniques he/she deems appropriate or effective, and decides on the amount of time to be spent on teaching the language aspects he/she considers important. Testing should resemble teaching in the sense that the same techniques that were employed during instruction should be used in the test, as this ensures students are tested as they were taught. Moreover, compatibility should also be ensured by testing only what was taught. Additionally, the number of tasks testing a particular language aspect should be proportional to the time spent on covering them in class. If vocabulary was given priority over grammar in a certain period of learning, then the number of tasks on the test should resemble this proportionality — the test designer should include more vocabulary than grammar tasks in parallel proportion to the time spent on these two aspects of language during instruction. Furthermore, it would be erroneous to utilize a certain teaching style in class and another in testing. For instance, a test will not yield reliable data if the teacher focuses on language knowledge more than on skills in his/her teaching practice, but then tests students' listening or writing ability by including considerably more of these tasks, or a disproportionately large amount of them in regard to elements of language knowledge, on the test itself. Additionally, a test should correspond to language use, which means that test designers need to include such test techniques that will reveal whether test takers possess a particular type of language knowledge or skill that is important for a relevant situation. For example, if students are tested for admission to a university and they are expected to already possess certain language knowledge and skills used for specific academic purposes, then test designers need to include tasks that resemble those situations which students will undoubtedly encounter in the academic context and in which this specific language knowledge or these particular skills are required.

Bachman and Palmer (2004) and Dimitrijević (1999) claim that *tests need to be designed in a way that encourages and enables test takers to perform at*

their best. To meet this demand, test designers need to refrain from: creating a test for the purpose of punishing students and thus from including content that has not been covered sufficiently or at all; using test techniques students are unfamiliar with; writing imprecise or misleading instructions for tasks; etc. Any such activity turns the test into an unreliable measuring instrument and raises the affective barrier that prevents students from realizing their full potential. Testing, just like teaching, will only likely be effective if the relationship between the teacher and students is based on respect. It is essential that students trust the teacher in the sense that the teacher provides them with accurate information, that he/she has a genuine interest in helping students learn, and that he/she uses effective teaching methods and conducts fair grading that allows for the making of unbiased decisions. Along the same lines, the teacher needs to be ensured that his/her students are willing to learn and are ready to invest time and effort in broadening their knowledge. Moreover, test performance is enhanced if students are adequately prepared for the test, i.e., if they are timely informed about the time and manner of the test to be administered and if they are familiarized with the content to be measured, the type of tasks the test will contain, and other particulars that might impact their test performance, such as the purpose of the test, the time allowed, the scoring method, and the like. To succeed in its purpose, the administered test needs to contain the same type of tasks students encountered in preparing for the test with the teacher, since giving them tasks they are unfamiliar with requires them to spend valuable time simply on understanding what they are supposed to do, time which would be better spent doing the tasks themselves or checking their answers.

A test needs to possess several key characteristics that ensure its quality. Bachman and Palmer (2004, pp. 17–18) refer to these characteristics as 'test usefulness,' a term that encompasses reliability, validity, authenticity, interactiveness, impact, and practicality. According to these authors, complementarity of these characteristics is needed, that is, a reasonable

balance among them needs to be found in each testing situation. Based on Bachman and Palmer's understanding and definition of the test's qualities (2004, pp. 17–43), a short explanation of each of them follows so that this point can be more fully understood:

- *Reliability*: consistency of measurement, or, in other words, it means that if the same measuring instrument is used to test the same or a similar population, similar results should be obtained;
- *Construct validity*: It is essential the test and each task it contains measure what they are intended to measure, so that the interpretations made on the basis of the results are meaningful and appropriate. To achieve this, tasks need to contain only the items measuring the intended knowledge or skill. In case a task includes an item which does not fall within the scope of the task's intended purpose, it is considered to be invalid;
- *Authenticity*: the correspondence of tests and/or test tasks to real-life situations which students may encounter. Tasks should simulate real-life situations in which test takers are likely to take part in real life and thus elicit the knowledge students would use in similar real situations;
- *Interactiveness*: the activation of the test takers' language skills, topical knowledge, and emotions in executing a task. The task is interactive if its execution requires the test taker's use of his/her foreign language capacity and topical knowledge and if it engages him/her emotionally;
- *Impact*: the effect a test has on the test taker, society, and the educational system. The test exerts an influence on different stakeholders as we always "use tests in the context of specific values and goals" (Bachman & Palmer, 2004, p. 30) and we make choices based on those;
- *Practicality*: the application of the test, i.e., how the test is implemented and whether its administration exceeds available resources.

Chapter 1

Topics for discussion

1. If you are a practicing teacher, how often do you create your own tests? How often do you use tests available to you? What are the most common sources of tests that you use?

2. Do you think that designing your own test is so complex that you feel discouraged to do it?

3. Do you think teacher-made tests have any advantages over ready--made tests? If so, what are they?

4. Do ready-made tests have any advantages when compared to teacher-made tests? If so, what are they?

5. What would help you in honing your test design skills?

2. TEACHER-MADE AND READY-MADE TESTS IN FOREIGN LANGUAGE TEACHING: PROS AND CONS[1]

As previously mentioned, foreign language teachers in Serbia and in other countries alike receive some instruction in foreign language assessment and a large number of authors are unanimous in claiming that practicing teachers need ample opportunities to enhance their assessment literacy skills in order to be able to create better quality tests and use them more knowingly. Burdened by many obligations, as well as due to their lack of theoretical and practical knowledge of test construction and a lack of professional development opportunities, a number of Serbian foreign language teachers extensively use ready-made tests. Such a practice might be justifiable to some extent. Namely, commercial tests are very likely to be devoid of language mistakes and be formatted well as well as to possess content validity, i.e., they virtually unmistakably include the material covered. Also, it goes without saying that ready-made tests free teachers from the often lengthy and at times daunting task of test design. They also represent a viable option for novice teachers lacking experience in test design. On the other hand, coursebooks typically come with just a few tests, for which reason assessment is likely to become a rather scarce pedagogical activity. Additionally, publisher-made tests typically do not provide a scoring system or corresponding guidance, which would seem to be necessary as a teacher may not be skillful in devising such a system on his/her own and interpreting the results. Along the same lines, ready-made tests often do not include tasks measuring students' language

[1] This chapter is a somewhat adapted version of the paper "The importance of teacher-made tests in foreign language teaching" published in 2017 in *Nasleđe*, *36*, pp. 285–296, in co-authorship with Vesna Pilipović.

skills (speaking, reading, writing, listening), but only knowledge (grammar, vocabulary, etc.). Hence, they too often fail to provide teachers with an overall picture of students' language knowledge. In the remaining part of this chapter, further comparisons are made between these two test types.

Due to a plethora of both subjective and objective reasons (e.g., heavy workload, busy schedule, lack of assessment literacy skills, lack of professional development opportunities, fear, etc.), many Serbian foreign language teachers use ready-made tests. However, such pronounced reliance on commercial tests bears clear negative consequences. Namely, relevant literature reveals that each test needs to be based on some pre-defined elements (e.g., purpose, definition of construct, etc. — see Section 2.2. for more information on these elements) which need to be defined by the teacher himself/herself as they depend on the students' characteristics, course requirements, contextual factors, inferences to be made, and the teacher's pedagogical beliefs. These elements determine the test's usability, as well as the validity of the data it yields. Since these elements are intrinsically not provided or addressed by ready-made tests, the purpose and usability of such tests are questionable, which further implies that the results acquired through them should not be used for making important pedagogical decisions.

Nevertheless, fortunately, relevant research indicates that the most prevalent foreign language assessment technique is still a test designed by a group of English language teachers, followed in frequency by those created individually, while those provided by a coursebook publisher are used third most regularly (Prošić-Santovac et al., 2019, p. 261). However, a dearth of studies have evidenced a lack of teacher assessment literacy skills (e.g., Marso & Piggie, 1988, 1991, 1993) and possibilities for their improvement (Glušac & Milić, 2021); hence, it remains imperative that teachers are given chances to enhance their assessment-related knowledge and skills and thus improve both their instruction and student learning.

As indicated by a number of authors (e.g., McMillan, 2000; Marso & Piggie, 1988), tests designed by practicing teachers usually contain errors, which then impacts both the test taking and the results. However, there are studies that confirm teachers' high reliance even on such flawed tests. Illustrative of this are the findings of Gullickon (1984, cited in Marso & Piggie, 1993), whose investigation in teachers' beliefs and attitudes towards testing revealed that teachers deemed their self-created tests "result in increased pupil effort, influence pupil self-concept, create desirable competition among students, improve interaction among pupils, improve the classroom learning environment, better focus teaching, provide a better learning experience for pupils, motivate pupil study, and accurately reveal pupil progress" (p. 153). Furthermore, some studies confirm that when assigning grades or judging students' progress in general, teachers place more importance on the results students obtain in classroom tests than on any other measuring instrument (Frey & Schmitt, 2007, p. 404). For that reason, teachers need to receive quality instruction and should be supported throughout their career in improving their test development, design, and interpretation skills.

Testing, as a means of monitoring students' progress and measuring the results of their learning, is an integral part of the teaching practice. It enables the teacher to check students' progress in learning, the efficacy of applied teaching methods and techniques, the achievement and achievability of set aims, etc. In addition, testing helps the teacher compare, group, monitor, and/or select students. Due to the reasons already established, tests may exert a minor, intermediate, or major influence on students, though they all share one common denominator: they help teachers make a number of pedagogical decisions. As pointed out by a number of authors (Alderson, 1999; Alderson et al., 1995, cited in Hidri, 2021), when a judgment of someone's learning is expressed as a number, that number will have no meaning in case it is unreliable and invalid. Therefore, it is of utmost importance a test be a precise measuring instrument in order to give the teacher accurate information based

on which he/she makes decisions. Needless to say, the teacher needs to be skilled both at designing and administering the test and interpreting its results.

Given that a significant number of foreign language teachers in Serbia are relying on the publisher's coursebook tests that accompany the coursebooks they use, the aim of this chapter is to emphasize the benefits and necessity of creating teacher-made tests by drawing on relevant literature on foreign language testing. By doing so, the author hopes to motivate educators working with future foreign language teachers to provide their students with relevant information and practical experience necessary for the creation of their own tests, as well as to encourage practicing teachers to embrace creating and applying tests on their own more willingly and frequently.

2.1. Concept, role, and importance of testing in the process of teaching

Language, as a complex form of human behavior, is composed of a number of interrelated and inseparable components: grammar, vocabulary, pronunciation, listening, writing, and speaking and reading ability, as well as of different competences, such as linguistic, sociolinguistic, and pragmatic competences and their sub-competences (Council of Europe, 2002, pp. 108–138). In order for students to achieve full mastery of a language and be able to use it independently, each of these elements needs to be given due attention in the teaching process. From time to time, it is important to assess how students are progressing with respect to all language components, what objectives have been achieved, how effective are the methods or techniques employed, whether it is possible to proceed to the next stage in learning, etc. Brown (2000) asserts that "[a] good teacher never ceases to assess students, whether those assessments are incidental or intended" (p. 402). However, available research shows that formative assessment is used less than summative (Frey & Schmitt, 2010; Assessment Reform Group, 2003, p. 12), most probably due to the reasons stated earlier in this chapter.

Since the teacher continuously makes pedagogical decisions of varying importance, it is essential that tests be based on the data gathered by means of a certain measurement procedure. Testing is only one means of conducting measurement, regardless of whether that measurement is done for the purposes of assessment or evaluation. In the context of the Serbian foreign language classroom, a test is undoubtedly the prevailing assessment/evaluation technique. Since the result of most assessments and evaluations is a grade that teachers, parents, or the school will use for making further decisions, a plethora of negative emotions are associated with the test as an instrument of measurement. However, a good test should help the teacher and his/her students in directing learning, instead of representing or being regarded as a threat or punishment.

Given that learning a language entails mastering a variety of its constituent elements, measurements of students' progress and knowledge need to be frequent and varied (Rudner & Schafer, 2002, p. 9; Shepard, 2000, pp. 44–48; McMillan, 2000, p. 3). Since the test development and design procedure is lengthy and presupposes the teacher's adequate knowledge and skills, it is not uncommon for foreign language teachers to use the tests that come with the coursebook they are employing. A number of studies (e.g., McMillan et al., 2002) indicate that there has been an increase in the use of commercial tests. While the reasons behind such a trend in practice should certainly be further explored, one thing is clear: teachers' awareness of the importance of their self-created measurement instruments needs to be raised. Moreover, teachers and student teachers alike should be equipped with practical skills for determining the quality of a test, as well as for designing their own tests. As discussed earlier, ready-made tests can be helpful only to a certain extent and it is teacher-made tests which are more useful measuring instruments, as they represent a much truer reflection of the process of learning through which students have gone (Shepard, 2000, p. 43), thus providing more relevant information. Moreover, as observed by Radić-Bojanić and Topalov (2016),

"not a single course book designed for the global market can perfectly match all the needs of a specific group of learners" (p. 141), but each could be used as a reference and a guide for steering the teaching/learning process in the direction that best suits a particular group of learners. What this also implies is that not a single test that accompanies such coursebooks can be suitable for every group of learners. Tests that contribute most to the improvement of learning are those designed by teachers themselves (Guskey, 2003, p. 6; Assessment Reform Group, 2003, p. 3).

2.2. Teacher-made vs. ready-made tests

Upon a review of contemporary, relevant literature on the design and qualities of teacher-made tests, a number of elements that ready-made tests do not possess have been identified. The absence of these elements in commercial tests speaks in favor of teacher-made tests as more effective, valid, and reliable measurements of classroom activity. The list of these elements includes the following:

(1) Test specifications

A number of authors (Bachman & Palmer, 2004; Brown, 2000; Bodrič, 2016; Hughes, 2003; Alderson et al., 2002; DiDonato et al., 2013) emphasize the importance of working out the details of the test before its actual realization commences. Test specifications include information such as the purpose of the test, the definition of the construct, test takers' characteristics, the test's structure, time allotment, the scoring system, and others; the elements Bachman and Palmer (2004, p. 87) include in the design and the operationalization stage. Similarly, Weir (2005) asserts that "[a] test should always be constructed on an explicit specification, which addresses both the cognitive and linguistic abilities involved in activities in the language use domain of interest, as well as the context in which these abilities are performed" (p. 14). The test designer should have all this

information in mind both while writing and grading the test, as well as during the analysis of the test's results and their interpretation, since the overall usefulness of the test depends on how the listed elements are conceptualized. In the same vein, Fives and DiDonatto-Barnes (2013) also argue that the best way to ensure the obtaining of reliable evidence that will be used for making credential decisions about students is the creation of a table of specifications, which should help the teacher "align objectives, instruction, and assessment" (p. 1).

Defining the purpose of the test should be closely tied to both the curriculum and syllabus and the planned outcomes of students' learning since teachers are obliged to cover the material these documents prescribe. Only the teacher knows what syllabus-required content he/she has covered with students, and how this process has been carried out (Bruce & Schmitt, 2010, p. 108). Implicitly then a test which comes along with a coursebook will almost surely fail to reflect and accurately measure the entire content prescribed by the syllabus, raising further doubts about the usefulness of such provided, ready-made tests. The most useful test is one that features and thereby measures the content the teacher has covered in the classroom and that he/she is obliged to cover. Additionally, many elements in the test writing process depend on the purpose of the test, including the choice and order of tasks (Dimitrijević, 1999, p. 105), the way in which answers are graded, and the interpretation of results (Bachman & Palmer, 2004, p. 96 — see Section 4.2. for more information and examples on the effect of the test purpose). Moreover, many test and task characteristics are dependent on how the purpose and the construct are defined, such as interactiveness, authenticity, reliability (Bachman & Palmer, 2004, p. 171–172), and validity (Frey & Shmitt, 2007, p. 416), etc. Despite specially trained professionals being typically included in both the writing of coursebooks and corresponding tests and generally adhering to the latest advancements in science in doing so, publishers' tests do not include test specifications, so it is questionable as to how clear the tests' purposes and aims are to a teacher utilizing them, as well as how he/she might then understand, interpret, and use the results.

Besides defining the purpose of the test, the definition of the construct is also necessary. Bachman and Palmer (2004, pp. 118–119) state that when defining the construct, it is possible to choose between a syllabus-based and a theory-based construct. While the former relates to all language elements contained within an instructional syllabus, the latter is grounded in "the theoretical model of language ability" (Bachman & Palmer, 2004, p. 118). This definition can, therefore, presuppose listing all the elements the teacher intends to measure or the components put forward by the theory of language ability (Bachman & Palmer, 2004, p. 117). The test designer's choice of testing techniques, writing of test items, devising of the grading system, analysis, and his/her interpretation of results all depend on the definition of the construct. Since as a rule neither the construct definitions nor the purposes of tests are provided by coursebook publishers along with the tests, it is doubtful whether the teacher can comprehend and use the results of the provided tests appropriately. Similarly, Fulcher (2010) adds the following: "The scores on 'general' language tests are not necessarily built on constructs relevant to the decisions that need to be made in a specific context" (p. 101). Along the same lines, as a note of warning, Bachman and Palmer (2004, p. 116) emphasize that defining the construct(s), i.e., specifying abilities to be tested, is crucial for justifying the use of the test and its results, and making intended inferences. They also add that "[w]hat this also means is that the test developer cannot simply accept, without question, the construct labels that other test developers have used, as either corresponding to the construct to be measured, or as being appropriate for this particular testing situation" (2004, p. 116). Every test developer, therefore, needs to define the construct(s) himself/herself based on the inferences he/she wants to make.

Bachman and Palmer (2004, p. 11) also claim it is important to know the test takers' personal qualities and take them into consideration when planning and writing a test, especially those qualities that pertain to their topical knowledge and affective schemata, since they impact both language use

and test performance (Weir, 2005, p. 53). More precisely, the test should be designed in a way that improves the test taker's performance, not in a way that hampers it (Bachman & Palmer 2004, pp. 12, 66).

(2) Testing sample

Publishers' tests that come along with coursebooks typically presuppose measuring the acquisition of knowledge covered in the few units preceding each test. However, since no single coursebook can meet all the needs of teachers and students who use it (Radić-Bojanić & Topalov, 2016), and assuming that a coursebook should not serve as the principal learning material, but, rather, as a supplementary source of information, the question is raised as to whether the teacher will really have covered the entire breadth of content appearing on and intended for measure by a publisher's test. One of the basic principles of testing is that we can test only what we have taught (Dimitrijević, 1999, p. 54), or, in other words, the test should include only the language that has been taught and used in class (Heaton, 1990, p. 12), for which reason the content that is tested should include only what the students have had a chance to learn. Otherwise, certain test qualities could be violated, such as content validity (Zhang & Bury-Stock, 2003), and we would probably obtain misleading results based on which erroneous decisions about students' progress would likely be made.

Furthermore, Heaton (1990, p. 12) suggests that in language tests for students sharing a common native language we should also include elements that account for the extent of interference between the two languages. Ready-made tests available through a coursebook cannot include items accounting for such interference as these books are intended for the global rather than a local market. Additionally, a number of authors (e.g., Heaton, 1990, p. 13; Dimitrijević, 1999, p. 138; Fives & DiDonatto-Barnes, 2013, p. 4) advise that when choosing the size and scope of the test sample, the test maker should first determine the percentage of time spent on covering certain language aspects

in class and then decide on the number of tasks so that it is in proportion to the extent to which a certain content was covered through instruction. Fives and DiDonatto-Barnes (2013) add to this the following: "Things that were discussed longer or in greater detail should appear in greater proportion on your test. This approach is particularly important for subject areas that teach a range of topics across a range of cognitive levels" (p. 4). Similarly, Dimitrijević (1999, p. 54), Heaton (1990, p. 13), Bachman and Palmer (2004, p. 13), and Bruce and Schmitt (2010, p. 108), also deem it necessary for a test to be a true reflection of both the teaching/learning that has occurred in the classroom and the material that has been covered. Accordingly, a logical conclusion ensues: a ready-made test cannot be a close reflection of the extent and depth to which specific parts of the content have been addressed through instruction; for this reason, test items in a ready-made test are almost surely bound to be ascribed a different value than they are given in a real classroom.

Contemporary coursebooks are conceived in such a way that attention is paid to individual types of language knowledge (e.g., grammar, vocabulary, etc.), language skills (e.g., listening, speaking, etc.), and competences (e.g., pragmatic, communicative, etc.). However, the tests that come along with these coursebooks generally do not include tasks measuring all individual elements of one's language capacity, but only some of them, so the question of their ability to measure full language mastery remains unanswered. Illustrative of this tendency is that tests including tasks aimed at measuring students' writing or speaking ability barely exist among ready-made tests available to teachers, so it is doubtful whether a teacher who relies heavily on these tests ever employs measuring instruments (beyond the publisher's tests) to check the elements not accounted for in these tests. Relevant literature suggests that teachers should design and employ batteries of tests (Dimitrijević 1999, p. 106) throughout the process of instruction, and those batteries should encompass tests measuring different components of foreign language knowledge. Only by conducting such tests can teachers gain insight

into students' overall language ability. Publishers' tests almost exclusively rely on a one-sided approach to testing, i.e., they are nearly always comprised of the same type or number of tasks, which is not in congruence with the very nature of foreign language learning, in which different elements of a language are taught differently (Dimitrijević 1999, pp. 28–29).

(3) Testing techniques

Firstly, testing techniques need to be aligned with the purpose of testing (as well as with the definition of the construct(s)) and they should contribute to its achievement (Dimitrijević, 1999, p. 223). Not only is the test expected to reflect the way teaching/learning has occurred in class and the material covered, but the tasks we decide to include in a test need to be familiar to the test takers (Brown, 2000, p. 410; Weir, 2005, p. 54), since the method we use to test them can affect their performance (Alderson et al., 2002, p. 44; Dimitrijević, 1999, p. 223). For instance, if a test includes a task whose purpose and type students are not acquainted with, it can discourage them, slow them down, or demotivate them to proceed with the test. Also, such an act decreases face validity, which is of great significance for the test results (Hughes, 2003, p. 33; Brown, 2000, pp. 409–410). Shepard (2000, p. 49) adds to this that a good test needs to include such tasks that aid students in realizing their full potential. For all the reasons stated, it logically follows that ready-made tests should not include tasks that students are unfamiliar with since it would prevent them from realizing their full potential in a testing situation.

The choice of tasks to be included in a test also depends on the teaching practice and the teacher's pedagogical beliefs. Moreover, if a teacher utilizes a traditional teaching style and believes that knowing individual elements of language is more important than putting them into use, this teacher's test would likely contain isolated tasks, i.e., tasks testing individual elements of knowledge, a practice which does not truly determine students' ability to use

their knowledge in real communication, but rather their ability to reproduce what they have learned. On the opposite end of the spectrum, if a teacher favors the communicative approach and thus considers that all the elements of knowledge should be put into practical use, how would it be possible for such a teacher to measure his/her students' communicative ability using publishers' tests, since these tests typically do not include tasks measuring speaking or writing ability? Therefore, not only should the test be in concert with the material covered in class and the manner in which it has been covered, but it also should mirror the teacher's pedagogical beliefs pertaining to the nature of learning, as well as his/her teaching style. Additionally, Bruce and Schmitt (2010, p. 108) warn that the use of examples on a test that have been derived from a book, entire commercial tests, or tests made by someone else who teaches or has taught the same subject opens up a number of validity questions.

Furthermore, the type of test and the tasks it includes influence the way students will prepare for the test (Rudner & Schafer, 2002, p. 8). If a student knows he/she will need to show the ability to synthesize information, the learner will prepare differently than in a situation when he/she knows the test will include multiple-choice tasks. Teacher-made tests should also include tasks that promote different levels of cognitive processing (Fives & DiDonatto-Barnes, 2013; Zhang & Burry-Stock, 2003; Shepard, 2000) (see also Chapter 3). In principle, the levels of cognitive reasoning in the test should be the same levels that students have been required to engage in during instruction. This does not imply that in publishers' tests there are no tasks promoting different levels of students' reasoning skills, but only the teacher working with a specific group of students knows at what levels of cognitive processing his/her students can operate, and it is only this teacher who can knowledgeably decide on the levels of cognitive capacity of the tasks a test should include.

(4) Test construction

When constructing a test, a test designer typically writes the test items, decides on the order of tasks, makes decisions regarding the context, writes task instructions, etc.

Writing test items and deciding on their order are greatly dependent on the test specifications (see point (1) in this section). They are also dependent on the test designer's knowledge of the test takers (Dimitrijević, 1999, p. 105). In other words, the items in a test should measure only the knowledge that students have had a chance to acquire in class, they should be pertinent to and informative for the intended test takers, and they should be written in language with which the test takers are familiar. Only a teacher who works with a specific group of learners knows what those learners find difficult, easy, or interesting, and what their cultural background is (Alderson et al., 2002, p. 40). How approachable a test is for test takers depends on all these components. Additionally, the test's length (number of tasks and items in each task) is "a professional decision made by the teacher based on the number of objectives in the unit, his/her understanding of the students, the class time allocated for testing, and the importance of the assessment" (Fives & DiDonatto-Barnes, 2013, p. 4). For all these reasons, only a teacher working with a certain group of students can know all the relevant details and thus practically take them into account when designing a test.

Language is best tested in a context (Dimitrijević, 1999, p. 59), but the context students are provided affects their test performance (Weir, 1993). In this light, not a single ready-made test can truly provide a context suitable for all the students doing a particular test. Only a teacher working with a certain group of students can choose an appropriate context for a specific group of learners since the teacher knows what his/her students find interesting and what will boost their test motivation (Weir, 2005, p. 53).

The number of tasks a test may feature should also depend on the test's purpose (Dimitrijević, 1999, p. 106), and only the teacher can determine the degree of difficulty of individual tasks, as well as their order (to range from the easiest to the most difficult), since they depend on the characteristics of individual students. A good test needs to include tasks of varying degrees of difficulty as only such a test can ensure sensitivity as an important test quality (Dimitrijević, 1999, p. 106). Moreover, the order of tasks a good test includes depends on the characteristics of the students the test is intended for, since what one group of students finds easy, another may find difficult. In this respect, the same task may be placed at the beginning of a test or somewhere close its end depending on what the test takers are like. A good test will also indicate to the teacher the areas of knowledge students find problematic (Heaton, 1990, p. 10). In summation, a ready-made test that comes along with a coursebook cannot suit all students in terms of the degree of difficulty of the included tasks or the order of those tasks.

The way instructions are worded also impacts how students will do the tasks in a test (Weir 2005, p. 57) (see Chapter 4). They also depend on how familiar students are with the tasks the instructions accompany (Bachman & Palmer, 2004, p. 190). Only a teacher working with a specific group of students can know such things. Also, the vocabulary used for writing instructions and task items needs to be completely familiar to students; otherwise, an unfamiliar word or a phrase may be detrimental in the sense that it can hamper the execution of a task, while almost certainly compromising the test's validity (Dimitrijević, 1999, p. 74).

(5) Scoring

In the majority of cases, information or guidance on scoring is not included in ready-made tests. However, a good test should include this information since the test taker's understanding of correctness affects the execution of

a task (Bachman & Palmer, 2004, p. 189; Weir, 2005, p. 63), which can be additionally motivating for the test taker. Moreover, as Shepard (2000) puts it, "The features of excellent performance should be so transparent that students can learn to evaluate their own work in the same way that their teachers would" (p. 60). In this regard, any scoring criteria that might come along with a ready-made test would significantly alleviate the burden teachers already carry, since without knowing the purpose of a test, or the definition of its construct, the teacher will likely face difficulty when trying to determine the scoring criteria for the test he/she is provided. More precisely, only the abilities and elements of knowledge that are included in a test's construct definition should be scored (Bachman & Palmer, 2004, p. 194) and without being informed about the purpose or construct(s) of a provided test, teachers cannot successfully devise the scoring system on their own. All the abilities and elements of knowledge not included in the defined purpose or construct may be sub-skills or secondary skills and should be disregarded if they are not part of the test's specifications. The information on what is intended to be measured by a particular test can thus only be obtained by consulting the test's specifications, but these are rarely offered together with publishers' tests. A test designer might think that the purpose of a task is obvious, but a teacher using the test could employ it for quite a different purpose. Using an inadequate task for measurement compromises validity and yields unreliable data that should not be used for making significant pedagogical decisions. In truth, no decisions should be made on the basis of such results.

(6) Interpretation and analysis of test results

If a teacher is unfamiliar with the purpose and construct(s) of a test, he/she will not be able to interpret the results it yields since the teacher does not truly know what the test measures. If we intend to get insight into our students' language ability, then this ability first needs to be defined as precisely as possible (Bachman & Palmer, 2004, p. 66), in the initial phase of a test's

development. If the definition of the construct is not provided or if the teacher has not devised it, the results obtained cannot be interpreted correctly, for which reason the construct validity is jeopardized (Bachman & Palmer, 2004, p. 21). The incorrect interpretations of test results can lead to faulty judgments and decisions that impact students in ways neither desired nor intended.

A good test has a positive rebound effect on students' learning and the teacher's teaching practice (Rudner & Schaffer, 2002, pp. 8–9) in the sense that answers to some questions can reveal the cause of a mistake, which can signal to the teacher where additional attention should be paid regarding a particular aspect of knowledge. The test can, therefore, include material that students have found problematic and worked hard to master. Ready-made tests can never include such tasks.

2.3. Conclusion

Studies conducted in different countries (Alderson et al., 2002; Bruce & Schmitt, 2010; Fives & DiDonatto-Barnes, 2013; McMillan et al., 2002; DiDonatto et al., 2013; Shepard, 2000) have all yielded similar results: teachers find it difficult to construct a test since test design and development entails a number of different types of theoretical and practical knowledge, and it is essential that teachers constantly hone their testing skills or develop tests in teams (Hughes, 2003, p. 58). Despite numerous difficulties teachers encounter in the test construction process and despite the flaws a teacher-made test can suffer from (Assessment Reform Group, 2003; Frey et al., 2005; Martinez et al., 2009), when compared to ready-made tests available to teachers, the tests they create on their own provide them with a more reliable picture of the learning process and its results. As such, the teacher-made test is an invaluable means of improving both teaching and learning.

Without a doubt, the tests that come along with coursebooks have several advantages in comparison to teacher-made tests: they include authentic language, have likely undergone some metrical checks (e.g., validation), they are generally formatted well and devoid of language mistakes, etc. For these reasons, they can be usefully applied as quick tests aimed at diagnosing areas where additional learning should occur before teacher-made tests are employed as progress or achievement checks. In order to create a truly good test, however, the following elements are necessary: the knowledge of the test takers; familiarization with the process of learning and learning objectives; a clear idea of the purpose of testing and inferences to be made; awareness of the context in which learning took place. Ready-made tests cannot account for these factors, indicating the potentially significant damage they can exert — the results they yield are not valid and can have far-reaching consequences for test takers. Unless the validity of the inferences made on the basis of the results generated by a test can be proven, they should absolutely not be relied on when making decisions about individuals (Bachman & Palmer 2004, p. 95). This is certainly the case regarding the results of commercial tests.

Chapter 2

Topics for discussion

1. What standardized tests have you taken so far, e.g., FCE, CAE, TOEFL, an entrance examination, etc.?

2. Were you in any way affected by the decision made based on the results of such a test?

3. Were you allowed to ask for additional information/help during such test taking?

4. Were you given a chance to improve your test score?

5. How did you feel taking such an exam?

⁂ ⁂ ⁂

6. What is/was the prevailing assessment technique in your EFL classroom?

7. What are/were classroom tests used for, e.g., grading, checking progress, punishing students, etc.?

8. Do/Did you have a chance to improve test scores? If you are a practicing teacher, do you give your students a chance to improve their test scores?

9. How do/did you feel taking classroom tests?

3. COGNITIVE PROCESSING IN TEACHER-MADE TESTS OF ENGLISH AS A FOREIGN LANGUAGE[2]

Cognitive processing is in essence the thinking protocol, reasoning, or the capacity of an individual to cognitively operate with certain information or knowledge. It may represent a general person's cognitive capacity, while it can also relate to a specific domain. Therefore, it can be viewed both as a general and a domain-specific ability. As an illustration, a particular individual may have highly developed reasoning skills in general, yet only engage in lower levels of cognitive processing with respect to a new skill he/she is acquiring in a particular domain. However, this person's general cognitive processing skills, in the illustrated case, will help him/her progress in the specific domain more quickly.

As suggested by Bloom et. al (1956), there are six levels of difficulty of cognitive processing: (1) knowledge, (2) understanding, (3) application, (4) analysis, (5) synthesis, and (6) evaluation. To illustrate, when a person memorizes something, e.g., a rule, without really being able to understand

---

[2] This chapter is a somewhat adapted version of the following papers:

"Analysis of English language test tasks for fifth- and sixth-graders in Serbia according to Bloom's Taxonomy" published in 2020 in *Inovacije u nastavi*, *33*(2), pp. 128–139, in co-authorship with Isidora Wattles and Nataša Marčićev;

And

"Analysis of English language test tasks for seventh- and eighth-graders in Serbia according to Bloom's Taxonomy" published in 2019 in *Nastava i vaspitanje*, *68*(1), pp. 35–50, in co-authorship with Vesna Pilipović and Nataša Marčićev.

or use it freely, he/she operates at the knowledge level. In this case, the person is only able to parrot back or regurgitate the information he/she has obtained. When the person is able to go beyond the mere memorizing of information and show understanding of it as well, he/she operates at the level of understanding. The three highest levels are commonly referred to as critical thinking (CT). They presuppose one's independent and creative use of information or knowledge in new situations. These three levels are typically the objectives that most educational systems worldwide strive to achieve: to enable their students to use the knowledge they acquire independently, for self-expression, and in novel ways.

Critical thinking is an important skill both for academic success and for thriving in today's world. The globalized society and Information Age we live in have created a demand for people who are skilled at managing and manipulating large pools of information. This calls for people's ability to discern between important and unimportant content or true and false data, to synthesize information, evaluate the trustworthiness of sources, make tenable decisions, and even to create something new and unique out of available resources. However, Halpern (1998) believes that there is solid evidence to claim that "many adults consistently engage in flawed thinking" (p. 449). Given that critical thinking is a component part of one's functional literacy (Glušac, Pilipović, & Milić, 2020), it is beyond doubt that a lack of critical thinking skills directly impacts the level of an individual's functional literacy. For this reason, critical thinking has become an educational priority at all levels of education in many countries, including Serbia, whose rulebooks on the syllabi and curricula for different elementary and secondary school grades rightfully list this ability as one of the goals of education.

Even though the ability to think critically emerges, and begins to be cultivated, before formal education commences, school must have the function of further developing and honing this skill in its students. What is more, as noted by

Halpern (1998), “Numerous studies have shown that critical thinking, defined as the deliberate use of skills and strategies that increase the probability of a desired outcome, can be learned in ways that promote transfer to novel contexts” (p. 449). More often than not, however, teacher education programs do not evolve at the same pace at which the world changes, so many of them still do not presuppose equipping future teachers with the knowledge and skills that are necessary for teaching CT. When they are employed, novice teachers seldom have an opportunity to gain relevant knowledge of CT through professional development programs, though they are still expected to promote and achieve it as a set educational objective. Too often, the meaning of the concept and the ways in which it can be taught and promoted are left to teachers to discover on their own. Yet to truly fulfill their professional obligations, they must somehow reach a full understanding of the concept and find appropriate ways to instill the associated skills in their students. Continuing with this logic, in order for teachers to discover whether their students have acquired these skills and whether any improvement is still needed, they would need to conduct assessments that would require their students to perform tasks at different levels of cognitive capacity. If teachers are not instructed in how to instill the needed CT skills in their students and monitor their development, they remain unprepared to test these skills properly and direct their further progress.

In light of this backdrop, the aim of this chapter is to expound a thorough theoretical framework related to cognitive processing in general and in foreign language teaching specifically. Additionally, the aim is to present the results of two studies that investigated what levels of cognitive capacity are required for completing tasks included in English language tests for students of the fifth through eighth grades of elementary schools in Serbia. More precisely, the primary purpose of these studies was to discover whether English language teachers in Serbia had been incorporating tasks requiring different levels of cognitive capacity from their students in the tests they had been designing.

The studies also sought to determine whether teacher-made tests assessing students' knowledge of English as a foreign language contained tasks at the higher-order thinking levels, which would imply students' ability to use the language independently and for communicative purposes, which is the ultimate goal of foreign language teaching/learning.

3.1. Understanding critical thinking

The educational objectives of any school system should comprise the development of students' affective, psychomotor, and cognitive domain (Bloom et al., 1956) in a stepwise fashion, progressing from simple to more complex behaviors. Bloom and his associates (1956) proposed a taxonomy of those objectives related to the three domains of development, specifying how students would be expected to be changed by the educative process (Bloom et al., 1956, p. 26). By doing so, they attempted to facilitate communication and understanding of the outcomes among different individuals and institutions responsible for designing or achieving them. The affective domain, as described by Bloom et al. (1956) includes such objectives that "describe changes in interest, attitudes, and values, and the development of appreciations and adequate adjustments" (p. 7). The authors admit that the objectives related to this domain were difficult for them to define, let alone for teachers to achieve. The psychomotor domain relates to the manipulative and motor-skill area, while the cognitive field refers to the development of intellectual abilities and skills. More often than not, an educational system neglects the achievement of at least one of these objectives, yet all are of vital importance for ensuring sound, comprehensive education that has positive, long-term benefits for its students.

It is the cognitive domain that is the centerpiece of this chapter. From the perspective of education, it presupposes different mental behaviors or cognitive

processes that students perform when and after learning either to understand or memorize content or to utilize the acquired knowledge in different situations and for various purposes. These behaviors encompass simple intellectual actions, such as: (1) knowing, or memorizing, things, facts, rules, paradigms, etc.; (2) understanding, or being able to transform, interpret, paraphrase, etc.; and (3) applying — putting into use in novel situations something a person has learned. More complex cognitive processes — (4) analysis, (5) synthesis, and (6) evaluation — comprise critical thinking. Each higher level is built on a solid basis of all the preceding levels of cognitive reasoning and reflects one's independent thought. Analysis, for example, relates to one's ability to break a whole into its constituent parts so as to analyze them. Synthesis is the ability to utilize the knowledge one has gathered to create something new, while evaluation is reflected in making purposeful judgments and presenting them. In this book, the six levels of cognitive processing will be referred to as Bloom's Taxonomy.

What specific mental actions an individual can perform at each of these six cognitive levels is probably best represented through a number of action verbs depicting the complexity of cognitive processes an individual is capable of performing at each stage. In other words, the action verbs commonly associated with each level of the taxonomy illustrate what mental activities a person can perform if he/she is to be considered capable of operating at a certain level of cognitive capacity. Consider the following table of action verbs associated with the six stages of cognitive processing.

Table 1. Verbs related to different levels of the cognitive domain

Level of Bloom's Taxonomy	What behavior can the person exhibit?	Action verbs
(1) Knowledge	Find or recall information	define, draw, duplicate, identify, label, list, match, name, outline, recall, recognize, select, show
(2) Comprehension/ Understanding	Construct meaning from given material	associate, classify, compare, comprehend, demonstrate, describe, differentiate, discuss, distinguish, estimate, explain, identify, indicate, interpret, relate, restate, select, summarize, translate
(3) Application	Use information in new situations	calculate, change, classify, compute, employ, execute, illustrate, implement, map, model, modify, organize, practice, present, show, solve, use, write
(4) Analysis	Make connections among ideas	break down, categorize, compare, context, contrast, differentiate, distinguish, experiment, illustrate, predict, question, research, separate, simplify, subdivide
(5) Synthesis	Produce something new/ original	compose, construct, create, criticize, design, develop, direct, formulate, generate, produce, revise
(6) Evaluation	Value information or ideas	argue, assess, conclude, convince, estimate, evaluate, grade, justify, measure, rank, rate, score, select, support, test

The first level presupposes storing information and one's ability to retrieve it and serves as a basis for all other ends or purposes of education (Bloom et al.,

1956, p. 33). In other words, the first level serves as the stage of memorizing information, facts, rules, etc., or learning the basics, in order to enable the understanding and then the application of the information at some later time. The subsequent levels are, thus, cumulative in that "the objectives in one class are likely to make use of and be built on the behaviors found in the preceding classes" (Bloom et al., 1956, p. 18). Hence, the taxonomy explains the progression from simple to complex behaviors, from the concrete or tangible to the abstract or intangible (Bloom et al., 1956, p. 30), from lower-order cognitive processes (knowledge, understanding, application) to higher-order cognitive processes (analysis, synthesis, evaluation). The former group presupposes simple mental operations with something tangible and concrete, while the latter presupposes an individual's capacity to deal with abstractions and create something new. Moreover, the latter is also considered to form one's capacity to think critically (Kennedy et al., 1991, as cited in Lai, 2011, p. 8), since students must perform several complex cognitive processes to deal with a novel situation.

It is exceptionally hard to provide a comprehensive and precise definition of critical thinking since it is composed of many skills and sub-skills, comprises different levels of complexity, and involves numerous personal traits. Critical thinking includes the cognitive and the non-cognitive domain. The former is commonly defined by the evaluation of various intellectual products (ideas, beliefs, etc.) in order to determine their qualities, such as relevance, validity, grounding in evidence, etc. (Pešić, 2011, p. 7). Authors such as Cohen et al. (2002) and Halpern (2003) add to this the metacognitive aspect of the cognitive domain, claiming it is just as important for an individual to evaluate the product as it is to evaluate the very process of thinking. More precisely, a person needs to be capable of monitoring his/her own thinking process in order to correct misconceptions, notice sources of potential conceptual mistakes, etc. These cognitive actions presuppose a number of cognitive processes, such as analysis, interpretation, drawing conclusions, and the like. In addition to evaluation and

metacognition as forms of critical thinking cognitive capacities, CT encompasses a number of non-cognitive qualities. These include dispositions or habits of mind (Facione, 1990), such as open- and fair-mindedness, inquisitiveness, flexibility, a desire to be well-informed, and the like (Lai, 2011). Additionally, it is paramount for a critical thinker to possess a number of personal traits such as systematicity, perseverance, tolerance, activism, social responsibility, etc. (Mirkov & Stokanić, 2015, p. 26). This clearly indicates that as one develops the ability to think critically, one also develops as a person.

Though there are certain discrepancies in their definitions of CT, owing to its rather complex nature and the contrasting approaches to it (e.g., philosophical, cognitivist), different authors agree that CT yields benefits beyond academic success. It represents an exceptionally important life skill without which it is impossible to thrive in today's world. In this regard, CT may be referred to as a global skill, but its application presupposes a thorough understanding of the domain in which it is to be applied (Cohen et al., 2002; Halpern, 2003). In addition, this skill can be viewed within the confines of the classroom — as a content-dependent skill, which is dependent upon the type of reasoning typical of a specific discipline (Glaser, 1984; McPeck, 1981; Paul & Elder, 2008). Cognitive processes do not evolve completely naturally, nor are they simply gained as one grows up; rather, they need to be taught carefully and practiced continually from an early age. In the school context, it is necessary for any subject teacher to foster the development of cognitive processes (Paul & Elder, 2008, p. 88) and use various teaching and assessment techniques that stimulate them. If this approach is adopted, knowledge ceases to be memorized and simply regurgitated; instead, its acquisition is gradual, as the student is engaged in a number of cognitive activities that help him/her first subsume the new material within the already existing knowledge base and finally use it in novel situations (Anderson et al., 2001).

As mentioned earlier, CT is undoubtably necessary for both academic success and thriving in today's world. However, what remains an unresolved issue is whether it needs to be taught as a separate school subject or within different subjects as a content-specific skill (Morais et al., 2019, p. 224). Here, CT is analyzed as a content-specific skill viewed within the confines of the English language classroom. As such, it can be utilized to instigate the learning of the foreign language and prompt its independent and creative use, which is the ultimate goal of language learning and, at the same time, represents the highest levels of Bloom's Taxonomy. Furthermore, CT ensures one's personal and professional success as it requires an individual's ability to approach information critically and to manipulate it successfully, independently, and creatively.

The need to foster different levels of cognitive processing and cumulatively improve students' capacities to think critically has been recognized by teachers in Serbia and abroad alike. Research studies investigating the teaching of CT within different school subjects and at different levels of education have begun to be carried out with the aim of exploring the effectiveness of applied teaching methods and techniques and highlighting areas for further development. In one such study conducted in Serbia that included 1,441 primary school teachers (Mirkov & Stokanić, 2015), the teachers were found to be aware of the need to promote students' CT and to be willing to do so. However, when correlating their attitudes towards teaching CT and their actual classroom activities, it became evident that they did not implement activities that promoted CT as much as they believed they should have. Regardless of teachers' readiness to teach CT, in a study reported by Mirkov and Gutvajn (2014), 856 eighth-graders from Serbia expressed their dissatisfaction with opportunities to foster their CT skills in school. They reported a lack of opportunities to ask questions, participate in discussions, or express their opinions. Similar results were obtained in a Portuguese study (Morais et al., 2019), in which despite university teachers expressing their willingness to promote CT within their own courses, the findings revealed that the teachers did not possess a complete

understanding of the CT concept, though they did strive to teach it using a variety of activities and learning materials. The study also showed that teachers encountered a number of obstacles, ranging from organizational (lack of time, group sizes, etc.) to institutional (lack of institutional culture and agreement on core principles/terms). Viewed solely within the context of English language teaching, a study conducted by Glušac and Pilipović (2016) indicated that primary and secondary school teachers in Serbia attempt to improve their students' CT by engaging them in Socratic questioning, a teaching/learning technique that requires students to investigate the nature and rationale of their thinking. The authors emphasized that this technique is beneficial in that "students are active participants in the teaching/learning process, as well as that they are responsible for constructing their own knowledge" (Glušac & Pilipović, 2016, p. 412). However, even though Socratic questioning is applied at the primary and secondary level alike, its true functionality remains doubtful and it is evident that some types of questions are used more than others (Glušac & Pilipović, 2016, p. 413). In light of all of this, familiarizing teachers with the notion of CT and its teaching principles should be a global necessity, so as to maximize its teaching potential. Needless to say, institutional support and adequate resources are highly crucial as well. Even more so, teachers need to be instructed in how to conduct assessment and learn whether they truly instigate CT, as the results of such assessments would likely point to areas that require improvement in terms of teaching and learning alike. In Section 3.3, two more relevant studies conducted in Serbia are presented.

3.2. Critical thinking and English language teaching and assessment

Foreign language learning lends itself well to teaching and improving cognitive reasoning. It is organized in a stepwise fashion and typically begins with the acquisition of isolated words, phrases, rules, and paradigms (knowledge), based on which a learner can understand another person's speech or writing

(understand), and only then be able to put a few memorized words or phrases and rules into practice (application). With the acquisition of knowledge, the learner becomes aware of differences between various linguistic options and their functions in different contexts (analysis), becomes capable of producing unique communication (synthesis) (Bloom et al., 1956, pp. 163, 169), which is the ultimate goal of foreign language learning, and develops the ability to perform different evaluations in accordance with either external or internal criteria or standards (evaluation). Moreover, CT is commonly associated with creative, analytic, and heuristic thinking, as well as with problem solving (Wattles, 2016, p. 6; Mirkov & Stokanić, 2015, p. 26). Not only is the thinking protocol teachable at the macro level (when considering the general process of foreign language learning), but it is applicable in everyday classroom situations (micro level). For instance, when teaching grammar or vocabulary, the teacher may prompt different levels of cognitive capacity of his/her students depending on the activity assigned. A case in point is a vocabulary exercise given in the form of a story from which some words have been omitted. For each of the gaps the student is offered a few possible solutions and he/she needs to select the most appropriate one. This activity is exemplary of the stage of understanding, as students display the ability to comprehend the story and complete it by selecting appropriate words. The same activity can be done in such a way that, instead of being offered possible answers, students need to provide their own solutions to complete the story. Such an activity is typical of the stage of application since students are required to use all their relevant linguistic knowledge acquired up to that moment and apply it in novel situations. Moreover, in the foreign language classroom students inherently encounter different cultures and lifestyles and are hence given a chance to break possible stereotypes and become open-minded, culture-sensitive, tolerant to differences, etc., all of which contribute to the development of important personal traits and dispositions that pave the way to successful CT.

In the English language classroom, as a content-dependent domain, it is important, and as previously seen, obligatory, to improve students' ability

to operate with and use the knowledge of the foreign language at different levels of cognitive complexity. In alignment with the prevailing approach to language teaching nowadays — the communicative approach, the ultimate goal of foreign language learning is the independent use of the foreign language by a learner in real-life, unrehearsed situations (Brown, 2000, p. 43). In the same vein, the general prescribed objectives for foreign language learning to which all European countries are committed outline a general progression in foreign language learning from using the language in strictly controlled, familiar contexts (level of understanding or application), to using it in less familiar ones (level of application or analysis), and on to usage in totally new ones (level of synthesis or evaluation) (Council of Europe, 2002, p. 24). This indicates the comprehensive support for an approach that gradually develops the skills of the students to utilize the gathered linguistic input. Thus, the teaching of cognitive processes in the domain of the foreign language classroom can be applied successfully to learning and/or improving language skills, such as reading (Wilson, 2016) and speaking (Rubin, 2017), since both require a gradual progression from controlled activities (at the application or analysis level) to free ones (synthesis or evaluation level). Moreover, even when acquiring language knowledge, such as vocabulary or grammar, a student can improve or build different cognitive capacities, as illustrated in the previous paragraph. Whenever students are faced with a language problem, i.e., a task to solve, they are prompted to use their domain-specific cognitive processes.

In Serbia, the planned outcomes of foreign language education for grades five through eight of elementary school (Rulebook on the Syllabus for the Second Cycle of Primary School Education and the Curriculum for the Fifth Grade of Primary School, 2016; Rulebook on the Syllabus for the Second Cycle of Primary School Education and the Curriculum for the Sixth Grade of Primary School, 2017; Rulebook on the Curriculum for the Seventh Grade of Primary School, 2018; Rulebook on the Curriculum for the Eighth Grade of Primary

School, 2018) clearly put forward educational outcomes which presuppose the engagement of different cognitive processes (e.g., comprehending, retelling, interpreting, describing, creating, expressing, etc.), the highest including the independent use of the foreign language for personal and creative purposes (e.g., taking part in communication while adhering to the sociocultural norms of the language, expressing one's own needs and interests, etc.) and dealing with novel situations, expressing and arguing one's point of view, etc. The planned outcomes for the four grades differ primarily in terms of the aspects of language knowledge to be acquired, rather than in relation to students' ability to use the language for executing tasks at different levels of cognitive complexity. It is intended that students develop and utilize the same cognitive processes in and across the four grades using level-appropriate language. This indicates that the syllabi and curricula in effect require the development of students' domain-specific cognitive processing, and critical thinking specifically, as a natural route of foreign language learning.

The combination of cognitive processing development and foreign language learning is beneficial for many reasons: it leads to the gradual acquisition of knowledge, which is more easily subsumed into the existing knowledge base; it is retained far longer than material learned through rote learning; it increases the general critical thinking capacity of students, as they can transfer the critical thinking pattern to other domains; it can boost students' motivation, as they are active participants and their opinions are valued; it provides better chances for the application of the acquired knowledge; and it resembles real-life situations and thus equips students with those abilities and skills they will need in their everyday living.

In order to understand fully what each level of cognitive processing represents and what foreign language activities may be done at each stage, consider the following list of activities, which is a modification of a list proposed by Bobrowski (2006).

Table 2. Modified model of the teaching/learning process of cognitive processes in the foreign language classroom proposed by Bobrowski (2006)

Level of Bloom's Taxonomy	Description	Key words	Example questions	Example language activities
Knowledge	Recalls information, definitions, descriptions, facts; Can cite or recognize accurate information regarding a question; Has some sense of what information is relevant.	Who, what, where, when, which; Find, choose, define, list, label, show, spell, match, name, tell, recall, select, organize, outline.	What is …? Where is …? When did …? Can you recall …? Can you select …?	Complete the sentences with the appropriate form of the verb *to be* in the Present Simple tense.
Understanding	Understands a concept, process, context etc.; Can process answers to critically—inquisitive questions and articulate what remains unclear; Has some understanding of how a certain item of knowledge is linked with other items in the knowledge base.	How, why; Relate, compare, contrast, illustrate, translate, infer, demonstrate, summarize, interpret, show, explain, classify, select, rephrase, distinguish, order, compare and contrast.	How did … happen? How would you describe …? What does it mean to …? Can you explain what is happening …?	Complete the sentences with one of the following words provided.

Application	Can apply and transfer particular knowledge to new situations; Can teach this knowledge to others.	Apply, construct, make use of, plan, build, develop, model, interview, experiment with, identify.	How would you use …? What would result if …? What elements would you choose to change …?	Complete the sentences with an appropriate present tense.
Analysis	Can solve complex problems by applying and generalizing concepts; Can produce a solution that is reusable and transferable to similar solutions.	Analyze, dissect, inspect, divide, simplify, solve, test, examine.	What inferences can you make? What conclusions can you draw? What would happen if …?	Indicate which word does not belong to each of the following groups of words and explain why.
Synthesis	Synthesis of the acquired knowledge and production of something unique.	Analyze, dissect, inspect, divide, simplify, solve, test, examine.	Can you suggest solutions …? Can you illustrate/describe …?	Describe one of the two people in the pictures.
Evaluation	Makes new linkages among concepts and problem solutions which have not been seen before; Makes judgments about the value of something.	Theorize, design, formulate, discover, make up, hypothesize, prove, invent, create an original work.	How feasible is the plan to …? Can you predict the outcome if …? What is necessary to discover …?	Write an essay containing 250 to 500 words, describing and evaluating the poem presented. In your description you should employ such terms as will reveal your recognition of the formal characteristics of the poem.

				Your principles of evaluation should be made clear, although they should not be deliberately described or defended. (Taken from Bloom et al., 1956, p. 198)

As was already pointed out in this chapter and as is obvious from Table 1, the progression of cognitive activities ranges from very controlled activities that simply require regurgitation of memorized information (knowledge), through activities that give an individual some independence of thought (application and analysis), to total freedom in using foreign language knowledge (synthesis and evaluation), which assumes using the language freely, creatively, independently, and for self-expression. For more examples of English language tasks requiring different levels of students' cognitive processing, see Section 3.4.

Even though both Bloom et al. (1956) and Anderson et al. (2001) claim that the taxonomy can be applied both for setting teaching/learning and assessment goals, how English language teachers around the world make use of it is questionable (Beaumont, 2010, p. 1). However, drawing on the ideas of Dimitrijević (1999, pp. 54, 122, 225–226) and Hughes (2003, p. 1) that we should test only that which we have taught and according to the manner in which it was taught, it suffices to say that if we promote cognitive processing in our classroom, then we should correspondingly test it by giving students tasks that would require them to exhibit cognitive capabilities at different levels of complexity while solving language tasks. Unfortunately, literature regarding testing cognitive processes in the domain of the English language barely exists. One of the two directly relevant papers is that by Fahim et

al. (2010), who found that those students who performed well on a Watson-Glaser Critical Thinking Appraisal measuring general CT skills performed better on EFL tests as well, which might imply that solving English language tasks involves using one's thinking capabilities. The second paper, written by Yanning (2017), reports on positive correlations between students' critical thinking and writing scores in the second language classroom. The remaining part of this chapter presents the results of two studies conducted in the domain of EFL (English as a foreign language) teaching and testing.

3.3. Research results

As was mentioned at the outset of this chapter, in addition to presenting fundamental concepts related to cognitive processing and CT specifically, the aim of this chapter is also to present the findings of two studies based on a corpus that included teacher-made tests for English as a foreign language designed for elementary school students in the fifth through eighth grades in Serbia. The purpose of both of these studies was to discover what levels of cognitive processing were required of students when completing the tasks contained in relevant tests. More precisely, the aim was to discover how frequently tasks operating at higher levels of cognitive complexity were being used, which would be an indicator of students' free and creative use of the foreign language.

The methodology of both studies, including the participants, time, and analysis of the corpus, was the same, as is explained below.

3.3.1. Description of participants

For the purpose of the two studies, in 2017 the researchers contacted English language teachers across Vojvodina and asked them to share their self-created tests with the researchers. Twenty-eight teachers from 10 towns consented

to share their tests for assessing their students' knowledge of English. The participant teachers were aged 30-45, had between 2 and 23 years of teaching experience, and taught two grades. Each teacher shared two tests, one for each of the two grades he/she taught: fifth and sixth or seventh and eighth. Altogether, 14 tests were gathered for each of the four grades analyzed.

The obtained tests were achievement tests measuring the quality of accumulated knowledge after a certain period of learning. Here, quality of knowledge signifies the ability to use the language information acquired for solving language tasks that require different kinds of manipulation of that information, i.e., different cognitive processes. The researchers examined teacher-made tests which had been created for this purpose since it was supposed that they contained only content which had been previously covered in class (Dimitrijević, 1999, p. 68).

Also worthy of mention is that the fifth- and the sixth-graders had been learning English for about 5-6 years at the moment the study was conducted, whereas the seventh- and eighth-graders had been learning English for at least 7 or 8 years and had probably achieved sufficient language proficiency to enable them to use the language for communicative purposes in a variety of situations.

3.3.2. Procedure

Upon receiving the tests, the researchers took on the arduous task of individually determining the level of each task in all the tests according to Bloom's Taxonomy, thus ensuring researcher triangulation. When classifying the tasks according to the level of cognitive capacity required from the student for completing the tasks, the researchers closely followed the definitions and examples of the six levels of cognitive processing put forward by Bloom et al. (1956, pp. 62–197), as well as the guidance for the classification of test tasks proposed by Bloom

et al. (1956, pp. 45–59). Moreover, the following factors were considered in the course of classification: the number of years of the students' learning of the foreign language, their age, the task instructions, the learning context, and the prescribed learning objectives. Once a level was determined for each task, the researchers compared their ratings. In cases where discrepancies emerged, the researchers analyzed those tasks again and compared them against the relevant descriptions and definitions. Once a consensus was reached with respect to each individual task, the authors counted the number of tasks at each of the levels of Bloom's Taxonomy in order to ascertain the number of tasks at each level included in the analyzed tests for a certain grade.

To illustrate how the processes of determining the levels of cognitive capacity of the tasks was done, consider the following task and read the description of the procedure that follows it.

[Example 1][3] (Grade 6) **Correct the mistakes in the following sentences:**

Last night, Samantha have pizza for supper.
My pet lizard was died last month.
Yesterday, I spend two hours cleaning my living room.
This morning before coming to class, Jack eats two bowls of cereal.
What was happened to your leg?

Since the instruction does not specify the type of mistakes students should look for, they would need to use all their linguistic knowledge gathered up to that point to analyze each sentence by breaking it down into its constituent parts and recognizing which parts contained a mistake so that they could then correct it. Hence, the researchers classified this task as *analysis*. If the instruction had specified that students should correct mistakes related to the verb, for instance, it would have fallen in the category of *application*, since students would have needed to apply particular knowledge, that of verbs,

[3] Examples of all test tasks in this book are given verbatim.

to conclude what was wrong with the verb form. On the other hand, if the instruction had specified that the mistakes were related to the Past Simple Tense, the task would have then been classified as a *knowledge*-level task, as they would have been required to remember specific rules, definitions, etc.

In addition to the instructions, the researchers also took into consideration the age of the learners when performing task classification. Depending on the syllabi and prescribed learning objectives for different grades, tasks were classified differently. For instance, a task in the fifth grade asking students to write dates in words (see Example 2) was classified as *application*, since this required them to use a completely new rule of saying dates and apply it in novel situations. If this task had been given to older students, already well acquainted with pronouncing and writing numbers and dates, it would have been classified as a *knowledge*-level task.

[Example 2] (Grade 5) **Write the dates.**

5/10 _____
12/2 _____
7/6 _____
23/5 _____
18/9 _____

The analysis of the tasks also revealed that the teachers designed certain tasks for whose execution students needed to perform at two different levels of cognitive complexity. Such tasks were regarded as and placed in a separate category of results.

Consider the following example illustrating a task operating at two levels of the taxonomy (*understanding* – write the questions, *application* – write short answers).

[Example 3] (Grade 5) **Write the questions and short answers.**

does / she / how often / dream?
_____?

live / at / you / do / school?
_____?

they / play / football / do/ when?
_____? […]

Since the needed forms of all the words are given, the students' task is simply to understand their meaning and order them accordingly in a way that makes each string of words a meaningful sentence. For this reason, the writing of questions is classified as *understanding*. On the other hand, when writing short answers, students need to make use of their gathered language knowledge (spelling, vocabulary, word order, grammar) to provide short answers. For that reason, the second part of this activity, writing short answers, is labelled as *application*.

3.3.3. Results for fifth-graders and sixth-graders and analysis

Altogether, 14 tests comprising 59 one-level and 7 two-level tasks for the fifth grade were analyzed. The results presented in Table 3 reveal that the prevailing level of cognitive processing at which the tasks operated in this grade was knowledge (29 tasks), followed by understanding (16 tasks), and application (13 tasks). There was only one task requiring any higher-order cognitive process (synthesis). Also, the results show that there were seven tasks whose execution required two levels of cognitive processing.

Table 3. Levels of cognitive processing in tests for fifth-graders

Level of Bloom's Taxonomy	Number of tasks at this level	Examples of types of tasks
Knowledge	29	Match the words on the left with the words from the box. Make adverbs out of these adjectives. Write the Past Simple for these verbs.
Understanding	16	Circle the correct option in each sentence. Write SOME and ANY to complete the sentences. Write the words in the right order to get sentences.
Application	13	Look at the picture and complete the sentences to say where different objects are. Make questions with the words given.[4] Write the following dates in words.
Analysis	0	
Synthesis	1	Describe the interior of your home.
Evaluation	0	
Two-level tasks	7	Look at the picture and complete the words that indicate items of furniture (*knowledge*). Then write a few sentences to describe where those items are (*application*). Fill the gaps with the appropriate forms of the verb TO BE (*knowledge*) and then make those sentences negative and interrogative (*application*). Write questions with the words given (*understanding*) and their short answers (*application*).
	Total: 66	

When the results pertaining to individual teacher tests were analyzed (Table 4), it was obvious that they all included tasks at the lowest level, while the majority

[4] The needed forms of words were not given, but rather the students were required to think of and use the correct form and order of the words provided, as in *your mom / cook dinner / now*.

involved tasks at the subsequent two levels (understanding and application). Higher-order thinking skills had completely been left out, with the exception of test 13, which contained a task classified as synthesis. Moreover, the results reveal that most tests contained several tasks at the first three levels of the taxonomy (see tests 1, 3, 6, 8, and 11), while a somewhat smaller number of tasks required students' engagement solely at the first two levels of the taxonomy (see tests 2, 5, 7, and 13). Tests combining tasks that operated at the first and the third level were not uncommon either (see tests 4, 10, and 12). In those tests containing tasks that required the operation of application there was typically only one such task per test (e.g., tests 1, 3, 4, 8, 10, and 11), and application-level tasks were most commonly combined with knowledge-level tasks (see tests 2, 5, 7, 9, 10, 11, and 13). In tests 6, 12, and 14, however, a greater number of application-level tasks were found.

Table 4. Levels of cognitive processing in individual teacher's tests for the fifth grade

	Knowledge	Understanding	Application	Analysis	Synthesis	Evaluation	Two-level
Test 1	3	2	1				
Test 2	4	1					1
Test 3	2	3	1				
Test 4	3		1				
Test 5	2	1					1
Test 6	2	2	2				
Test 7	3	2					1
Test 8	1	1	1				
Test 9	2						1
Test 10	2		1				1
Test 11	1	2	1				1
Test 12	3		2				
Test 13	1	2			1		1
Test 14			3				
TOTAL	29	16	13	0	1	0	7

The analyzed tests for the sixth grade show a somewhat different picture (Table 5). Out of 66 tasks included in the analyzed tests, most were either classified as understanding (23 tasks) or application (21 tasks), followed by those at the first level (knowledge) of Bloom's Taxonomy (17 tasks). These tests also included five tasks at the higher levels of cognitive processing (analysis – 3 tasks, synthesis – 1, task and evaluation – 1 task). The tests for this grade did not include tasks whose performance required the use of two different levels of reasoning.

Table 5. Levels of cognitive processing in tests for sixth-graders

Level of Bloom's Taxonomy	Number of tasks at this level	Examples of types of tasks
Knowledge	17	Complete the following sentences by putting the verbs in brackets in the Present Perfect. Complete the following phrases with a suitable verb.[5] Complete the table with either a noun or an adjective missing.[6]
Understanding	23	Match the expressions with the pictures. Put the words in the correct order to make sentences.[7] Complete the dialogue with the words offered.
Application	21	Write advice for the following situations using SHOULD and SHOULDN'T. Complete the sentences with the passive voice in a suitable tense. Kim did a lot of things yesterday morning. Write a sentence for each picture.

[5] The verbs were offered to the students and they needed to recognize which of them collocates with each phrase given.

[6] This task measured students' knowledge of word formation. For the nouns offered, the students needed to supply their adjective forms and vice versa.

[7] The needed forms of words were given, students just needed to arrange the words so as to form a sentence.

Analysis	3	Complete the questions and answers.[8] Study the following pairs of sentences and decide which one is grammatically correct. Correct the mistakes in the following sentences.
Synthesis	1	Make true sentences about you using the following verbs and ideas.
Evaluation	1	Write down a thing you are not allowed to do and a thing you can do and explain why that is so.
Two-level tasks	0	
	Total: 66	

When the distribution of tasks at different levels of Bloom's Taxonomy is analyzed from the perspective of individual teacher tests (Table 6), it can be noticed that only three teachers did not include tasks at the lowest level of cognitive complexity in their tests (see tests 2, 9, and 14). Also, it is evident from the results that the tasks requiring complex cognitive processing (analysis, synthesis, and evaluation) were few and apart (e.g., tests 2, 6, 11, 12, and 13). As is evident from the table, most teachers included tasks at the first three levels of the taxonomy (see tests 1, 3, 5, 7, 8, 11, 12, and 13), some combined lower-level tasks with analysis (see tests 6, 11, and 12), while only very rarely did a teacher do so with other higher-order thinking levels (see tests 2 and 13).

[8] The verbs in the sentences can be in different tenses. The students needed to analyze the given part of the question so as to infer which tense should be used to complete the sentences.

Table 6. Levels of cognitive processing in individual teacher's tests for the sixth grade

	Knowledge	Understanding	Application	Analysis	Synthesis	Evaluation	Two-level tasks
Test 1	1	3	1				
Test 2		2	2			1	
Test 3	2	3	1				
Test 4	1	4					
Test 5	2	1	1				
Test 6	1	2		1			
Test 7	1	1	2				
Test 8	2	1	3				
Test 9		1	2				
Test 10	3		1				
Test 11	2	3	2	1			
Test 12	1	1	2	1			
Test 13	1	1	1		1		
Test 14			3				
TOTAL	17	23	21	3	1	1	0

3.3.4. What do the results for the tests administered to fifth-graders and sixth-graders show us?

Generally speaking, the presented results are quite unsettling since the vast majority of the tasks included in the analyzed English language tests do not fulfill the scope of the three levels comprising CT. In other words, the analyzed tests do not help students improve their domain-relevant CT skills, which implies that they are not given a chance to use the acquired language freely and creatively, but are only asked to reproduce it. In Table 4, for example, only one task (see test 13) from the fifth-grade tests is at a higher level of the taxonomy, which would require students to use the language for self-expression. In the sixth grade, the picture is only slightly better, as is evident

in Table 6 (see tests 2, 6, 11, 12, and 13), where five tasks are shown that would prompt an individual to creatively use his/her gathered knowledge. All the other tasks for both grades require only the application of the low-order thinking skills.

Along the same lines, it is further unsettling that the tasks included in the tests for the fifth grade belong to a great degree to the lowest level of the taxonomy, asking students simply to remember/recall/regurgitate stored information. Even though cognitive reasoning is indeed cumulative in nature (Bloom et al., 1956, p. 18), i.e., that for the performance of cognitive activities at all levels of complexity a person needs to know the relevant rules, definitions, and paradigms, and that there are certainly justifications for the teaching of knowledge, as pointed out by Bloom et al. (1956, pp. 32–36), the teaching and testing of a foreign language should not be solely based on separate language items. Students need to be exposed to a variety of situations in which they would use the acquired knowledge for communicative purposes. Yet as an illustration of the absence of such an approach, in the analyzed tests for the fifth grade there was found only one such task (see Table 3, level of synthesis), asking students to describe the interior of their home. However, the analysis of the tasks for the same grade reveals that teachers did combine two levels of cognitive reasoning in certain tasks (see Table 3, two-level tasks), most commonly the knowledge and application levels. On the one hand, such tasks should prove useful for both students and teachers, as they require the application of knowledge students have previously shown they possess. On the other hand, such tasks might not be in concert with recommended test construction practices. Namely, when discussing multiple choice constructions, Dimitrijević (1999, p. 95) warns against those questions whose execution directly impacts the execution of subsequent tasks. The same warning might apply to other test techniques as well, since if students make a mistake or fail to do one test item, they inevitably fail to do the following one(s), just as is illustrated in Example 3 in Section 3.3.2. The situation

regarding students' exposure to activities allowing more freedom was shown to be only slightly better in the sixth grade.

The teachers' seeming insistence on declarative or receptive knowledge in the analyzed tests is also contrary to what is prescribed by the rulebooks for the two analyzed grades (Rulebook on the Syllabus for the Second Cycle of Primary School Education and the Curriculum for the Fifth Grade of Primary School; Rulebook on the Syllabus for the Second Cycle of Primary School Education and the Curriculum for the Sixth Grade of Primary School). These documents clearly indicate that students need to possess both receptive and productive types of language knowledge and to be able to communicate both in written and oral form. However, the analyzed tests show a clear inclination towards declarative or receptive knowledge despite the students for whom the tests had been designed having been learning English for at least 4 or 4.5 years at the moment of testing and, supposedly, possessing enough language knowledge to be able to use it freely and creatively, at least to some extent.

Since the analyzed tests for the two grades were constructed by the same teachers, the analysis of the results of the levels of cognitive capacity required in English language tests presented in Tables 3 and 5 reveals that the participant teachers implemented tasks at different levels of the taxonomy for the two grades. More precisely, the majority of the tasks found in the tests for the sixth-graders required understanding and application, whereas for the fifth-graders the tasks were shown to operate at the first two levels. Such a finding is encouraging to a degree as it indicates the teachers' awareness of the increased cognitive capacities of their older students. Also, the finding that a greater number of the tasks were determined as falling within the scope of understanding is aligned with the claim of Bloom and his associates (1956, p. 89) and Wattles (2016, p. 159) that understanding is the most prevailing intellectual level both in school and college. At the same time, however, the same results for the sixth grade are discouraging, since only 5 out of 66 tasks

in the analyzed tests are at levels which call for the free and creative use of the language. If the test design applied in the analyzed tests is indeed a mirror reflection of the teachers' general approach to testing, then this finding most probably indicates that the participant teachers employ such teaching and testing techniques that focus almost exclusively on separate items of the language system, rather than integrating those individual items into some form of cohesive whole. Such a practice would then be contrary to what is prescribed by the Rulebook on the Syllabus for the Second Cycle of Primary School Education and the Curriculum for the Sixth Grade of Primary School, which clearly emphasizes students' use of the language and prescribes that operative tasks should be more complex than for the previous grade. Moreover, the analysis of the results of individual teacher tests presented in Tables 4 and 6 indicates that the teachers most commonly combine tasks on the first and the second or the second and the third level of the taxonomy for fifth- and sixth-graders, respectively. In instances where tasks operating at higher levels were included, they were always combined with understanding and application tasks (see Table 6, tests 2, 11, 12, and 13), or knowledge and understanding tasks (see Table 4, test 13 and Table 6, test 6), while there was only one such higher-order thinking task per test.

3.3.5. Results and analysis for seventh-graders and eighth-graders

Fourteen tests were analyzed for the seventh grade, comprising 60 tasks altogether. As presented in Table 7, the prevailing levels at which these tasks operated were understanding (32 tasks), followed by application (14 tasks), and knowledge (10 tasks). Only a few tasks could be classified as demanding higher-order cognitive processes (3 – analysis and 1 – synthesis).

Table 7. Levels of cognitive processing in tests for seventh-graders

Level of Bloom's Taxonomy	Number of tasks at this level	Examples of types of tasks
Knowledge	10	Match the words on the left with their synonyms on the right. Complete the sentences with an appropriate form of *be going to*. Translate the following words into English.
Understanding	32	Read the text and answer the questions in full sentences. Complete the sentences with one of the following words. Complete the dialogue using *someone*, *somewhere*, *something* etc.
Application	14	Complete the sentences using the passive voice in an appropriate tense. Complete the following sentences with an appropriate past tense. Write sentences using the given words.[9]
Analysis	3	Name 3 things that make you happy, three that stress you and three things parents should do to keep their children healthy. Indicate the word that does not belong to each group and explain why it does not belong. Indicate which of the following sentences are in the Present Simple.
Synthesis	1	Describe one of the two people in the pictures - what he/she looks like and is like. Complete the sentences with your own ideas.
Evaluation	0	
	Total: 60	

A closer look at individual teachers' tests (Table 8) reveals that there were teachers who completely excluded tasks at the first level from their tests (see

[9] The form of the words offered needs to be changed so as to be able to form a sentence.

tests 3, 4, 8, 9, 11, 13, and 14). Rather, their tests contained only understanding-level tasks (see tests 4 and 13), understanding and application tasks (see tests 3 and 11), application tasks only (see test 14), or understanding and synthesis (see test 13), while two tests contained tasks at rather challenging levels of the taxonomy (see tests 8 and 9) – application and analysis, of which test 9 also included tasks at the understanding level. In the tests which included knowledge-level tasks, the tasks were combined with those at the level of understanding (see test 1), understanding and analysis (see test 6) or, more typically, with those at the level of both understanding and application (see tests 2, 5, 7, 10, and 12). One test combining tasks at the levels of knowledge, understanding, and application also included a task operating at the level of analysis (see test 6).

Table 8. Levels of cognitive processing in individual teacher's tests for the seventh grade

	Knowledge	Understanding	Application	Analysis	Synthesis	Evaluation
Test 1	2	3				
Test 2	1	2	1			
Test 3		4	1			
Test 4		6				
Test 5	1	1	2			
Test 6	2	2		1		
Test 7	2	2	1			
Test 8			3	1		
Test 9		2	1	1		
Test 10	1	3	1			
Test 11		3	2			
Test 12	1	1	1			
Test 13		3			1	
Test 14			1			
TOTAL	10	32	14	3	1	0

The analyzed tests for eight-graders included 14 tests comprising 66 tasks altogether. The results given in Table 9 show that these tests were composed almost exclusively of tasks at the two lowest levels of the taxonomy (23 tasks at the knowledge level and 31 at the level of understanding). Only four tasks belonged to the higher-order levels (3 – analysis and 1 – synthesis).

Table 9. Levels of cognitive processing in tests for eighth-graders

Level of Bloom's Taxonomy	Number of tasks at this level	Examples of types of tasks
Knowledge	23	Circle the correct form of the verb *to be* in each passive sentence. Complete the sentences with *was/were +ing*. Give examples of the following geographical items that are preceded by *the*.
Understanding	31	Order the sentences to form a dialogue. Paraphrase the sentences using the words in brackets. Match the words and their definitions.
Application	8	Complete the sentences with the Past Continuous or Past Simple tense. Complete the sentences with *a* or *the* where necessary. If the article is not needed, put *x*. Make a question for each of the following sentences so that the bolded word is the answer.
Analysis	3	Are the following characteristics good or bad? Circle the word that does not belong to the group according a certain criterion. Read the text and the descriptions of people and decide which person is the best candidate for each position.
Synthesis	1	Complete the sentences with your own ideas.
Evaluation	0	
	Total: 66	

The analysis of individual teachers' tests for eighth-graders (Table 10) shows that all but three teachers (tests 4, 13, and 14) employed elements at the knowledge level. Teachers mainly combined tasks which function at the first two levels (see tests 3, 5, 6, 7, and 9), while, besides these two levels, seven teachers also incorporated tasks functioning at the application level (see tests 1, 2, 4, 10, 11, 12, and 13). When such elements were used, only one such task was present per test, with the exception of test 2 (containing two tasks at the level of application). This is in contrast to the practice these same teachers displayed in the tests for seventh-graders, in which it was not uncommon to find two or three application-level tasks in a single test (see Table 8, tests 5, 8, and 11).

Table 10. Levels of cognitive processing in individual teachers' tests for the eighth grade

	Knowledge	Understanding	Application	Analysis	Synthesis	Evaluation
Test 1	1	3	1	1		
Test 2	3		2	1		
Test 3	3	1				
Test 4		3	1			
Test 5	1	3				
Test 6	1	3				
Test 7	2	4				
Test 8	4			1		
Test 9	3	2				
Test 10	1	2	1			
Test 11	3	2	1			
Test 12	1	3	1			
Test 13		3	1		1	
Test 14		2				
TOTAL	23	31	8	3	1	0

3.3.6. What do the results for the tests administered to seventh-graders and eighth-graders show us?

The findings are surprising in that the same teachers were found to have given more lower-order cognitive tasks to older students. In other words, far more tasks at the lowest level of Bloom's Taxonomy were given in the eighth grade than in the seventh grade. Moreover, a smaller number of application-level tasks were given to eighth-graders than to seventh-graders, whereas the numbers of understanding, analysis, and synthesis tasks were nearly identical. This is indicative of a general trend among the teachers to give eighth-graders tasks that simply require their recall of information. Such findings lead to the conclusion that the tasks given to eighth-graders for the purpose of assessing their knowledge of English are cognitively easier than those given to seventh-graders. The knowledge of rules, definitions, and paradigms in a language is indisputably the basis for performing more complex thinking protocols, and it is beyond doubt that there are a number of justifications for the teaching of knowledge (Bloom et al., 1956, pp. 32–36), but such levels of cognitive processing should not be predominant either in language teaching or testing (as is the case with the analyzed tests) since such a practice decreases students' communicative competence and deprives them of the opportunity to use language creatively. This is especially true in cases involving students who have been learning a language for a number of years and therefore would be expected to function as relatively autonomous language users.

Generally speaking, these findings are also in accordance with what Bloom et al. (1956, p. 89) claimed to be a trend both in schools and at colleges — the emphasis on those intellectual abilities and skills that involve comprehension. However, since testing should be a reflection of the teaching process (Dimitrijević, 1999, pp. 54, 122, 225–226; Heaton, 1990, p. 13; Hughes, 2003, p. 1) and the tasks the test is composed of should be those types of tasks practiced during the learning process (Dimitrijević, 1999,

p. 122), the findings are discouraging, as the participant teachers displayed an inclination to prompt eighth-graders to simply regurgitate information (knowledge level) despite their cognitive and linguistic abilities almost surely being well beyond this level. Since the ultimate goal of language learning is communication, the usability of the knowledge displayed by the students of the participant teachers in communicative tasks or situations is doubtful. Tasks at higher levels require language production, not mere recognition and/or remembering, but were quite scarce in the obtained sample. Again, if the test situation is reflective of the teaching process, then it appears obvious that eighth-graders had been primarily taught rules and definitions and had not been given many opportunities to apply or use knowledge for communication. In contrast, seventh-graders seem to have been taught, at least to some extent, to more fully understand and apply knowledge rather than to simply recall isolated pieces. Moreover, an analysis of the individual tests (Tables 8 and 10) shows that some of the tests for the seventh-graders excluded tasks on the first level of thinking entirely, whereas several tests for the eighth-graders were composed predominantly of tasks at the lowest levels (Table 10, tests 3, 5, 6, 7, 9, and 14).

The findings indicate that the prevailing testing approach among the participants was the structuralist approach (Heaton, 1990, p. 15), implying a somewhat traditional teaching style. However, an important element must be taken into consideration here — courses in foreign language assessment at English language departments in Serbia have only recently been introduced; hence, many participant teachers may not have received any formal education regarding test construction. We must also refrain from jumping to the conclusion that their teaching is reflected in their tests, even if this should be the case, since they may not be aware of this priority. Moreover, the participant teachers may not have been acquainted, and almost certainly not to an adequate degree, with the notion of critical thinking or cognitive processes during their formal education and likely have only encountered few occasions

about it during their professional development. In this light, they should not be criticized significantly for not including tasks at different levels of cognitive complexity in their tests and should certainly be held more accountable for assessing their students' knowledge of rules and definitions than for assessing their free and creative use of the language, despite some of their students having studied English for at least 7 years (since the first grade of elementary school). A justification for such a decision may also lie in the types of tasks the teachers appeared to use in their tests being more objective and easier to score, whereas the tasks requiring knowledge production (typically those operating at the higher levels of Bloom's Taxonomy) may imply subjectivity on the part of the teacher in grading, something which they may have been seeking to avoid.

3.4. Conclusions and pedagogical implications

The two studies presented in this chapter mirrored a few other studies conducted during the 1980s, the essential difference between them being that those earlier studies had included the analysis of tests for all school subjects, not solely for a foreign language. What is common to all of them, however, is that they yielded similar results. For instance, in the study reported by Fleming and Chambers, (1983, cited in Marso and Piggie, 1991), the researchers analyzed 342 tests containing 8,800 tasks, also utilizing Bloom's Taxonomy. The findings indicated the following: the junior high school level teachers included tasks operating at the knowledge level in 94% of cases, while elementary and senior high teachers constructed their tests with 69% of tasks operating at the lowest level of the taxonomy. The analysis of the correspondence between the subject area and levels of cognitive processing revealed that math and science teachers included tasks at higher levels of cognitive processing, but that this was not a customary practice among teachers of other subjects. In the same vein, another study was conducted by Billeh

(1974, cited in Marso and Piggie, 1991) in Lebanon schools. It included the analysis of 33 science tests for students of seventh grade through tenth grade. The analysis revealed that 72% of the tasks were at the knowledge level, 21% at the level of comprehension, and a mere 1% at the application level. Tasks at levels of cognitive processing other than these three were not found. Beyond this, the author came to other interesting revelations: the levels of cognitive complexity varied in accordance with the teachers' training in test design, not with the grade they taught. Additionally, the author found that more experienced teachers tended to give more knowledge-level tasks. Conversely, in a study reported by Black (1980, cited in Marso and Piggie, 1991), it was found that the cognitive complexity of tasks did not correspond to the extent of the teacher's training. Having analyzed 48 science tests, the researcher found out that the tasks included in those tests did not reach beyond the level of application and that the level of complexity varied between science subjects. Furthermore, in their own study of teacher-made tests and the levels of cognitive complexity required by them, Marso and Piggie (1991) collected a corpus of 175 tests, both for science and social studies subjects. The tests comprised 6,504 tasks altogether and the researchers ascribed each of them a corresponding level of Bloom's Taxonomy. Key findings included the following: 72% of the tasks were determined to be at the level of knowledge, 11% at the level of comprehension, 15% at the application level, 1% at the level of analysis, and fewer than 1% at the two highest levels. When analyzed individually, most tests were found to include exclusively or predominantly tasks measuring at the knowledge level. The largest number of higher-order tasks was found in science tests. To this book's author's best knowledge, few other studies have been done that explore the levels of cognitive processing in tests, including tests of foreign languages.

As mentioned earlier in this chapter, fostering higher levels of cognitive reasoning cultivates students' ability to manipulate the information they acquire, as well as their independent and creative use of that information for

self-expression. If only lower levels of cognitive complexity are in the teaching focus, students may be limited to simply remembering the information, recognizing it, and applying it only in controlled/guided contexts. Given that we should test only that which we teach and how we teach it, it goes without saying that in order to be eligible to test different levels of cognitive processing, teachers must first adapt their teaching practice so that it includes activities promoting different levels of reasoning. Only when students become acquainted with such tasks as part of the teaching routine can their ability to perform at different levels of cognitive complexity be truly measured.

The results obtained by the two studies presented in this chapter indicate that the tests for the four analyzed grades did not include tasks that increased in complexity along with the intended progression of the respective grades. In other words, the levels of cognitive complexity demanded by the test tasks did not seem to methodically increase with age, cognitive maturity, or level of language proficiency, but instead seemed to be a rather random selection by the teachers. The results make it obvious that there is a tendency among English language teachers in Serbia to design tests that predominantly include low-level thinking tasks. Given that students' cognitive capacities and language proficiency should be well developed by the higher grades of primary school, it was surprising to discover that in the tests for the seventh and the eighth grade there was a paucity of tasks requiring higher levels of cognitive operation. In both of these grades, tasks at the second level of the taxonomy were most dominant, closely followed by those at the level of knowledge (eighth grade), or almost equally by those operating at the level of knowledge and application (seventh grade). The situation was shown to be slightly better in the tests for the sixth grade, in which the majority of the tasks were at the levels of understanding and application, whereas it appeared least favorable in the fifth grade, for which the majority of the tasks were at the knowledge level. All in all, what is evident from these examinations is that instead of an increase in levels of cognitive complexity with age and grade

level, there appeared to be a rather random selection of levels, not aligned with the students' probable cognitive maturity or linguistic proficiency. Few tasks operating at the higher levels of cognitive processing were found in the tests for all grades, indicating that the prevailing tendency among Serbian teachers of English as a foreign language is to tests students' ability to recall and reproduce the acquired information, rather than to use it communicatively, independently, creatively, and for self-expression.

The collected sample of tests was strongly indicative of an inclination among participant teachers to utilize a more or less similar approach to testing — one focused mainly on language elements and lower-order cognitive processes. This implies that the predominant testing approach in the analyzed tests was structuralist, favoring discrete-point testing, instead of integrative, which presupposes the communicative function of the language. This is a finding the researchers had not presumed they would obtain, yet which, despite its unfavorable implications, is a valuable revelation nonetheless. Among others, Heaton (1990) warns against such focus-on-form practice as it can "indeed have a harmful effect on the communicative teaching of the language" (p. 10). Moreover, as the participant teachers likely lacked a formal education in test design, they might have been designing tests that do not truly reflect their actual teaching practices. In other words, they could have been employing a teaching approach that was not structuralist, yet still relying on such an approach in testing. Should this assumption be proven correct, it would mean that they were violating one of the fundamental principles of testing — to assess in a manner that reflects the way the students were taught (see Section 1.1). Anderson et al. (2001, p. 254) warn against instructional and assessment activities not being aligned, since it decreases the instructional validity of the assessment and the likelihood of students' good performance on external tests.

If the obtained findings are indeed reflective of the manner of teaching, then the results remain rather unsettling for different reasons. In this light, they could

imply the participant teachers' lack of knowledge of critical thinking in general or of cognitive processes and how they can be tested, a situation which would demand that they be provided with opportunities to gain insight into these elements immediately, as success in both their further education and the job market, which their students would be bound to encounter, largely depends on one's thinking abilities, a situation already predominant for quite some time.

The gathered tests analyzed in the two studies may only constitute one measuring instrument contained in a battery of tests assessing different types of knowledge and skills. Hence, we must not jump to the conclusion that the participant teachers never require, or offer opportunities to, their students to use the language for communicative purposes. Further investigation into the origin of this situation would be beneficial and could reveal whether or not such results might also be attributed to a mismatch between the teachers' teaching and testing practice, one which they might be unaware of and which, as pointed out by Anderson et al. (2001), could be detrimental to successful test performance. Moreover, the results necessitate familiarizing Serbian English language teachers with the notion, teachability, and testing principles of CT, as well as informing them about the benefits and pitfalls of the predominant testing approach they seem to have adopted in order to ensure quality teaching of CT and quality foreign language testing.

3.5. Examples of English language tasks functioning at different levels of Bloom's Taxonomy

What follows is a collection of examples of tasks found in the analyzed tests measuring the knowledge of English as a foreign language given to students of the fifth through eighth grades in Serbia, whose results are presented in this chapter. The examples selected and presented here are aimed at illustrating different levels of cognitive complexity and helping the reader understand how the determination of task-level difficulty is done. Each example task is

followed by a short comment from the book's author regarding the assigned level of cognitive complexity.

LEVEL 1: Knowledge

[Example 4] (Grade 5) **Write the Past Simple form of these verbs.**

Order _____
Turn _____
Drop _____
Stay _____
Carry _____

The task in Example 4 simply requires the recall of isolated pieces of information.

[Example 5] (Grade 7) **Translate the words.**

sight	weather
rarely	brand
journal	geyser
simple	passport
look forward to	instead […]

Even though translation is listed as a cognitive activity typical of the level of comprehension, the task in Example 5 does not test translation, i.e., understanding, but simply the recall of the meaning of isolated lexemes.

LEVEL 2: Understanding

[Example 6] (Grade 5) **Complete the sentences with a comparative or superlative form of the adjective.**

Saturday is _____ (good) day of the week because I don't have to study then.
Novak Đoković is _____ (good) tennis player in the world.
Rihanna is a _____ (bad) singer than Beyonce.
Boiled vegetables _____ (bad) meal of all. It's so ugly!

The task in Example 6 requires that students show understanding of when a certain adjective form is used or required. It is not only the form that is tested, but the understanding of the context when a certain form is needed.

[Example 7] (Grade 6) **Complete the sentences with a suitable word.**

People have breakfast _____ they get up.
a) after that b) before c) after

Before I go to sleep, I get _____.
a) dressed b) undressed c) home

Smart people are interested _____ many things.
on c) at c) in […]

The task in Example 7 checks students' reading comprehension and ability to choose a word that best suits a particular context.

LEVEL 3: Application

[Example 8] (Grade 6) **Make questions to the answers.**

What time _____?
I think I got up around 5 o'clock.

What _____?
I heard someone shouting in the street.

Where _____?
Dave was sleeping in the bedroom.

The task in Example 8 calls for the application of the language knowledge students have gathered up to the moment of testing. Namely, they are required to use the gathered knowledge in a completely new situation here. Moreover, versatile knowledge is needed for the successful completion of this task: knowledge of different tenses, knowledge of forming questions, knowledge of word order, etc.

[Example 9] (Grade 7) **Fill in the gaps!**

It _____ (not, work). I think it's broken.
_____ (it, rain) at the moment?
I _____ (have, I) lunch in the cafeteria every day.
Sheila _____ (love) reading books. […]

In the task in Example 9 students are required to apply the knowledge they have acquired up to the moment of testing in novel situations. Also, as is the case with Example 8, versatile knowledge is needed for the successful completion of this task and students need to show that they possess, and that they can successfully decide which, particular knowledge to apply for completing each item.

LEVEL 4: Analysis

[Example 10] (Grade 6) **Correct the mistakes in the following sentences:**

Last night, Samantha have pizza for supper.
My pet lizard was died last month.
Yesterday, I spend two hours cleaning my living room.
This morning before coming to class, Jack eats two bowls of cereal.
What was happened to your leg?

Since the type of mistake is not indicated, in order to successfully complete Example 10, students need to break down and analyze all the sentences and determine where the mistake is in order to correct it.

[Example 11] (Grade 6) **Which sentences are in the Present Simple?**

Anne is always coming late.
First I get up and then I have breakfast.
I'm here.
She lives in Sofia.
Mandy's buying a pet dog as she's very lonely.

This task can also be categorized as analysis since answering the question requires students' analysis of the given sentences, especially of their verb

phrases, in order to arrive at a conclusion about the tense in which they are. As can be seen, the task is aimed at sixth-graders who would have recently learned the two tenses the sentences are in. Hence, success in this task requires them to break down the sentence and to analyze which of its parts compose a verb phrase and whether the verb phrase identified is in the required tense.

LEVEL 5: Synthesis

[Example 12] (Grade 5) **Use these words and expressions to make true sentences about yourself: *very good at*, *quite good at*, *not bad at*, *not very good at*.**

_____
_____
_____
_____

To complete the task given in Example 12, the students are required to synthesize all the knowledge they have acquired up to the moment when the test is administered to produce something new. Even more so, students are required to show a good command of the foreign language by constructing their own sentences, calling for self-expression, rather than providing rehearsed answers.

[Example 13] (Grade 6) **Describe your typical day.**

I usually wake up _____
_____
_____
_____

In Example 13, students are invited to construct their own response to the task by applying all the knowledge they have acquired up to the moment when they are tested. The task allows them to be independent and creative foreign language users.

LEVEL 6: Evaluation

[Example 14] (Grade 6) **Write down a thing you are not allowed to do and a thing you can do and explain why that is so.**

_____
_____
_____

In Example 14 not only are the students invited to construct their own response and thus show language independence and creativity, but they are also asked to evaluate something and support their opinion with valid arguments.

⁂

As pointed out earlier in this chapter (see Section 3.3.2), when determining the level of cognitive reasoning of a task, a number of criteria need to be taken into account, such as the age of students, their background knowledge of the language, the instruction for the task, and the task itself. Consider the following examples. Both are taken from the same test for seventh-graders.

[Example 15] (Grade 7) **Make sentences from the words given.**

missed / the / London / to / we / coach.
_____.

practice / you / USA / in / did / the / English?
_____?

with / us / tennis / didn't / play/ Sara.
_____. […]

[Example 16] (Grade 7) **Make sentences from the words given.**

we / try / to / speak / French / in / Paris
_____

they / decide / to / visit / the / museum

_____

I / make / a / sandwich / last / night

_____ […]

At the first glance, the two tasks might seem to test the same aspect of language knowledge — order of words in an English sentence. However, they operate at different levels of cognitive complexity, or, in other words, they require students to display different levels of cognitive capacity. Simply put, Example 16 is more difficult, or demanding, than Example 15 and requires more attention and thinking on the part of the test taker. Conversely, in Example 15 all the sentence elements are given in the very form that is needed to compose a grammatically correct sentence. Additionally, a punctuation mark is also given in Example 15, thereby suggesting the order of the words. This activity (Example 15) is at the *comprehension* level, as students need to infer the meaning of the sentence by studying the words in a string and indicate the correct order of the words that would make the string comprehensible and meaningful. On the other hand, the words in each string in Example 16 are not given in the form required to make grammatically correct sentences by simply ordering them. Here, in other words, not only do the test takers need to order the words in a way that makes a string a meaningful stretch of text, but they also need to adapt their forms so as to make them grammatically appropriate for a particular context. More precisely, the students need to use an appropriate form of the given verb so as to make the sentences both correct and meaningful. Furthermore, punctuation marks are not given, which makes the construction of the sentences yet more challenging, as students need to decide on the type of the sentence as well. For all the reasons stated, this task operates at the *application* level as students need to show their ability to utilize the knowledge acquired up to that point in novel situations.

How the classification of a task can be affected by students' grade level and background knowledge is illustrated by the following example.

[Example 17] (Grade 5) **Translate the following sentences.**

Ja nisam Elizabeta. _____
[I'm not Elizabeth.]

Mi nismo u bašti. _____
[We're not in the garden.]

Vili nije dobar u slikanju. _____
[Willy's not good at painting/drawing.]

Considering their grade and background knowledge, the task can be determined as operating at the *application* level. Namely, fifth-graders would need to apply different types of language knowledge in a novel situation, i.e., to form novel sentences utilizing the language knowledge acquired up to that point in their learning. More precisely, they need to remember specific vocabulary, grammar, spelling, and syntax rules in order to construct sentences they have not heard or seen before. This same task given to older students would be commonly classified as *understanding*, since its main goal would be the transfer of meaning, i.e., translation. Yet this task given to the same fifth-graders could be classified differently if the students doing it were only being required to remember original sentences in English from their textbook. To illustrate, suppose that the sentences in Serbian in Example 17 are the translations of the English sentences that the students encountered in certain units in their coursebook and the teacher intends to check if they remember those original sentences by asking them to recall them. In such a scenario the task would be classified as a *knowledge*-level task since the students would be required to simply recall the exact information they memorized.

Chapter 3

Topics for discussion

1. How familiar were you with Bloom's Taxonomy and its different levels of cognitive processing before reading this chapter?

2. If you are a practicing teacher, how often do you check what levels of cognitive complexity you have included in your test?

3. Do you think it is important to have tasks operating at different levels of complexity in every test?

4. Consider the following tasks and determine the levels of cognitive processing they require. Also, explain how you have arrived at your conclusion about the level, i.e., what factors you considered when deciding on the level of Bloom's Taxonomy.

(Grade 5) **Make adverbs out of these adjectives**.
Frank speaks _____ (loud).
I run _____ (slow).
Steve spells _____ (bad).
Bob sings very _____ (good).

(Grade 5) **Write SOME or ANY**.
He doesn't have _____ bananas.
Do you have _____ cheese?
They need _____ water.
Shone wants _____ potatoes.

(Grade 5) **Fill in the blanks with appropriate forms of TO BE.**
I _____ sleepy.
We _____ busy.
You _____ pretty.
You _____ all happy.
The children _____ good. […]

(Grade 5) **Write questions and short answers.**
Your mom / cook dinner / now?

_____

Yes, _____

Your friends / play football?

_____

Yes, _____

(Grade 6) **Use the future simple:**
I _____ (turn on) the fire.
A: She's late. B: Don't worry. She _____ (come).
The meeting _____ (take place) at 6 pm.
If you eat all the cake, you _____ (feel) sick.
They _____ (not be) at home at 10 pm.

(Grade 6) **Answer the questions using the expressions of frequency: *once*, *twice*, *three times*, *four times … + a day, a month*, *a week*, *a year*.**
How often do you use the internet?
How often do you go jogging?
How often do you play volleyball?
How often do you listen to music?

(Grade 6) **Make negative sentences.**
Cathy forgets her homework.
The dog is sitting under the tree.
He does magic tricks.
Sam and Max live with their mother.
You are having fun!

(Grade 6) **Make the future simple.**
A: There' someone at the door. B: I _____ it.
A: I'm moving house tomorrow. B: I _____ and help you.
I _____ there, I promise.

(Grade 7) **Translate the following words and phrases:**
trainer
elegant
skinny
smooth
cowboy

(Grade 7) **Match the phrases.**

1. That's the assistant	a) where the Queen lives.
2. Buckingham palace is the place	b) who cut Brad Pitt's hair.
3. That's the dog	c) where we go on holiday.
4. Cathy is the hairdresser	d) which belonged to Madonna.
5. I like the camp side	e) who served me in the shop.
6. He bought the guitar	f) which followed me home.

(Grade 7) **Choose the Present Perfect or Past Simple.**
I _____ (never/be) to Vienna.
My great grandfather _____ (have) five sisters.
He _____ (live) in Manila for a year when he was a student.
Oh, no! I _____ (lose) my wallet! […]

(Grade 7) **Fill in the gaps with the correct verb forms (First Conditional).**
If you climb that tree, I _____ (give) you an apple.
If I _____ (get) the prize, I will be delighted.
They _____ (not, be) late if they _____ (leave) on time.
If it _____ (not, rain) tomorrow, we _____ (go) to the beach.

(Grade 8) **Complete the sentences with WAS/WERE + -ING.**
What _____ you _____? (do)
Last Thursday William _____. (shop)
They _____ for a nice present. (look)
He _____ some cakes. (make) […]

(Grade 8) **Make the adjectives from the underlined nouns.**
The sun was shining.
There was too much noise at the party.
He has a lot of luck. […]

(Grade 8) **Underline the odd word out.**
pretty attractive good-looking plain
overweight fat quiet well-built
friendly boring reliable helpful […]

(Grade 8) **Make questions for the bolded part of the sentence.**
Insects eat **plants**.
Italy produces **good wines**.
The Germans dink a lot of beer.
Dolphins eat **small fish**.
They went **to Spain**. […]

5. After you have determined the levels of cognitive complexity the given tasks operate at, describe how easy or difficult it was for you to determine their levels. Why do you think this was so?

6. Look at the levels of cognitive complexity you determined for the given tasks and consider whether the levels would be changed if the tasks were given to lower and/or higher grades. If so, how?

7. Give a suggestion on how one of the tasks given above could be changed so that it would operate at one of the three highest levels of Bloom's Taxonomy.

4. TEST TASK INSTRUCTIONS[10]

Task instructions are an important component part of the test. They are intended to convey to the test taker how he/she is supposed to proceed in doing the task, what particular aspect of knowledge is tested, how correct answers will be scored, etc. Based on the type of the test, other information may also be included, such as the time allocated or suggested for doing each task, where to record answers if not on the same sheet on which tasks are given, etc. Moreover, as Bachman and Palmer (2004) observe, "Test instructions also serve as an important affective goal: motivating students to do their best" (p. 182). Instructions are a means of communication between the test designer and the test taker; hence, they need to be succinct, unambiguous, and precise. Any failure to provide the test taker with full information in the instructions regarding the test and its tasks may have adverse consequences on test performance, or as put forward by Boyle & Fisher (2007), "If the test user does not provide clear instructions, is unfamiliar with the test materials or procedures for administration, and does not encourage the test taker, then the results will not be an accurate reflection of the learner's level of ability or skill" (p. 20). Ineffective or poor instructions can impact the quality of the test as a measuring instrument, for which reason decisions made based on the results of such tests might be questionable.

Owing to the importance of the information they convey, instructions are regarded as a critical test component (Hughes, 2003). However, few authors

[10] This chapter is inspired by the research conducted by the book's author for the purpose of writing the paper "Quality of written instructions in teacher-made tests of English as a foreign language," published online in 2021 in *English Teaching and Learning* in co-authorship with Mira Milić.

writing about foreign language test construction address instructions. Such a paucity of guidelines on how to write effective test task instructions has been recognized by a number of authors (e.g., Lakin, 2014; Glušac & Milić, 2021), all of whom have agreed that instructions have been undeservedly overlooked and should be given due attention in textbooks on foreign language test development and design. In addressing instructions in foreign language tests, Glušac and Milić (2021) observed that among all the textbooks they surveyed not a single textbook used for teaching academic courses in foreign language assessment provided enough information on writing effective and quality test task instructions or illustrations of such instructions. Moreover, it is not an uncommon finding that some authors writing on foreign language test design fail to address instructions entirely (e.g., Fulcher & Davidson, 2007).

Different authors have contrasting views on the importance of instructions in classroom tests. For some authors, most likely those who do not cover instructions in their textbooks on test development and design (e.g., Fulcher & Davidson, 2007), instructions are seemingly not regarded as being of great significance in teacher-made tests since students are likely seen as growing accustomed to them during instruction or test preparation. These authors might believe that students do not really need to read instructions when taking tests as they would be assumed to already know what they are supposed to do and what is expected from them in each task. Others (e.g., Lakin, 2014; Weir, 2005) claim that instructions are indeed significant and that students should develop the habit of reading and following them carefully in order to avoid any adverse situations, especially when taking tests other than those teacher-made ones with whose features and procedures they are already thoroughly familiar.

According to Bachman and Palmer (2004, p. 182), who, as observed by Glušac and Milić (2021), address the topic of instructions in remarkably greater detail than other authors in the field of foreign language testing, there

are two types of instructions: general and specific. The former are typically given to students well before the test is taken or when it is administered and they relate to the nature of test taking. More precisely, these instructions relate to students' expected behavior during the test administration, available resources, etc. For instance, students need to know whether they can leave the examination venue during test administration for any reason, what resources they are allowed to make use of during the test, such as a mobile phone or a dictionary, whether there is any set order in which tasks are to be completed, etc. Additionally, general instructions can also inform students on the scoring system, if it is the same for the entire test, as well as on how and where they are expected to record their answers if the same rule applies to all the tasks. If the information regarding these aspects of the test is different for each task, then it is specified individually for separate tasks and is a specific instruction. General instructions can be given either orally or in writing. On the other hand, specific instructions pertain to each individual task included in the test and relate to what students are expected to do, how and where to record answers (if the procedure is different for different tasks), how the answers will be scored, etc.

Throughout this chapter a number of examples of test task instructions will be used to illustrate the points discussed. The examples are derived from a large corpus of teacher-made tests the author has referenced in recent years for writing papers on different aspects of tests created by English language teachers for assessing the knowledge of their students. When analyzing instructions in this chapter, the analysis can relate only to some aspects of an instruction in question, those that illustrate a point being discussed. The same instruction may be lacking in other aspects as well, but they might not be addressed in the same analysis as they are not illustrative of or relevant to the point(s) discussed.

4.1. Instructions and test qualities

The failure to write effective instructions, i.e., those that are succinct, unambiguous, and precise, can lead to the violation of different test qualities, such as objectivity, validity, reliability, and authenticity. In other words, the way test instructions are worded may affect the quality of the test, students' test performance, their test results, and, finally, the teacher's and other's decision making.

First of all, task instructions can violate *objectivity*. If a task instruction fails to provide test takers with complete information as to what they are expected to do, different students can provide different types or forms of answers, which in turn makes scoring difficult and can affect objectivity. To illustrate this, consider the following example[11] of a task instruction by trying to do the task according to the given instruction, as doing it will reveal how imprecise the instruction is.

[Example 18] (Grade 5) **Answer the questions.**

How old are you?
Have you got many toys?
Which months do people in Serbia go on holiday?
What is your favourite book?
What does your friend look like?
What do you like doing after school?
What time do you go to bed?

Different forms of answers are possible in this task. For instance, to answer the first question, one can write: *11*, *Eleven*, *I'm 11*, *I'm eleven*, *I'm 11 years old*. The teacher is then likely to be troubled as to how to grade these different answers, whether any should be rejected for not being given in the form of a full sentence, or if they should all be acceptable. In contemplating on how

[11] The examples of all test task instructions in this chapter are given verbatim.

to score tests, the teacher needs to ask himself/herself two crucial questions, the first of which is: Did I instruct the test takers as to what form of answer I am looking for? If not, then different forms of answers are all acceptable. Since the instruction for this task does not specify the type of the answer, then all forms of answers should be accepted. If it is important that students give a specific form of answer, then such information must be stated explicitly in the instruction. The second crucial question is that of the purpose of the task. In other words, the teacher needs to determine a specific purpose for each and every task in the test, as well as to define a relevant construct for each task, and write an effective instruction having both the determined purpose and construct in mind (see Section 2.2, point (1)). To illustrate this, we can ask ourselves what the exact purpose(s) of the task given in Example 18 is — to assess students' reading comprehension (whether they understand the questions by responding to them appropriately), students' communicative ability (whether they can provide adequate answers), students' language knowledge (whether their answers reveal the knowledge of grammar, vocabulary, syntax, etc.), or something else? We do not know for sure what the purpose of the given task is, or its construct, but, based on the instruction, which does not specify what form of answer is expected, we can suspect that language knowledge (i.e., the use of correct spelling and grammar constructions) is not important. In other words, based on the instruction, we do not get an impression that the teacher expects full sentences, so he/she is not interested in checking students' sentence construction or particular elements of language knowledge. He/She seems to be more interested in measuring students' communicative competence or reading comprehension, and thus any form of answer would seem to be allowed. If, for instance, the purpose of this task is to check grammar competence, then the instruction needs to specify that answers should be given in the form of full sentences, and could read: *Answer the questions in full sentences*. What is more, to prevent getting responses of varying length to questions such as this one, which would further complicate the grading process, the instruction should include additional

information regarding the desired length of the answer sought, just as it is discussed in Section 4.3, point (6).

How does this impinge objectivity? Based on the given instruction, in the task in Example 18, it would be erroneous to punish students for not providing full sentences as answers. If students were to be penalized for not providing full sentences, then objectivity would be violated, as this would signify that the teacher is interpreting the instruction the way he/she thinks is appropriate. Any such involvement on the part of the teacher, in the sense of interpretation, implies a degree of subjectivity.

To avoid subjectivity in grading students' answers, it is essential that the instruction clearly inform students how they are expected to perform. Moreover, the purpose of the task needs to be clear to the teacher since, based on the intended purpose and the defined construct for each task, he/she should create a key for these tasks predicting all possibly relevant answers his/her students might offer. It goes without saying that the key needs to be devised in accordance with the instruction. To illustrate this, in Example 18, for instance, all expected and acceptable forms of answers should be listed in the key, as the instruction clearly fails to specify whether any particular form is expected.

In addition to compromising objectivity, ineffective instructions can also violate *validity*. As an illustration, consider the following example task.

[Example 19] (Grade 7) **Translate the following adjectives into your mother tongue:**

amicable
popular
shyly
kindly
self-confidently
cleverly

The task here is not valid since the items do not test what is claimed to be tested in the instruction. More precisely, the instruction informs the test takers that the task tests their knowledge of adjectives and asks them to translate the given examples of that word class, but the given list of items includes only two adjectives (*amicable* and *popular*), while the other listed words are adverbs. The instruction is, thus, misleading and the task fails to test what it is supposed to test.

Test instructions can also compromise *reliability*. Namely, the instruction needs to be unambiguous so as to ensure it is understood in the same way upon each repeated taking of the test. Should it be worded unclearly, i.e., in a way that the test taker can interpret it differently upon each repeated taking of the same test, the test designer cannot expect to get the same or similar results by applying the same measuring instrument, which in turn means that reliability is violated. Consider the following example that illustrates an ambiguous instruction.

[Example 20] (Grade 6) **Write the names of the places.**

You get the money from here _____
You borrow books from here _____
Planes take off from here _____
You can see cows and chickens here _____
You get a meal here _____
People stay here on holiday _____

The instruction invites students to provide any word they know that denotes the places defined. More precisely, there are multiple answers for all the items in this task. For instance, you can get money from a number of places, including a bank, an ATM, or a post office. Given that there are a number of answers to each question, it is very likely that upon each repeated taking of the test including this task, one and the same student would provide different

answers, which makes this task not reliable. To improve its reliability, students could be referred to a particular unit in which the vocabulary tested in this task is covered. For instance, if the instruction read *Write the names of the places covered in Unit 3*, then reliability would be improved since students would be told explicitly which words they would be expected to provide and they would probably offer the same answers every time the test were to be repeated. The suggested improvement of this instruction would also lead to enhanced objectivity, since the answers provided by students would not be subject to the teacher's interpretation; specific answer (e.g., lexemes covered in Unit 3) would be sought and their correctness or appropriateness would not be dependent on the teacher's judgment.

Another test quality that can be affected by instructions is *authenticity*. As put forward by Bachman and Palmer (2004), this is a test quality that significantly contributes to the test's usefulness as it presupposes "the degree of correspondence of the characteristics of a given language test task to the features of a TLU[12] task" (p. 23). In other words, tasks are considered to be authentic if they resemble real-life situations and if students are likely to find themselves in the same or similar situations in real life as the ones in which they are invited to engage in on a test. Students' answers to authentic test tasks thus help us generalize beyond their test performance (Bachman & Palmer, 2004, p. 24). Given the educational objectives established for the primary school level foreign language instruction in the context of the grades surveyed in Serbia to which the example tasks correspond, the TLU domain is the use of language in everyday situations that are relevant to the student. Consider the following example as an illustration of tasks that can be characterized as being relatively high in authenticity.

[12] TLU stands for target language use domain.

[Example 21] (Grade 6) **Make questions for the answers given.**

What time _____ ?
I think I got up at 5 o'clock.

What _____ ?
I hear someone shouting in the street.

Where _____ ?
Dave was sleeping in his bedroom.

The task is relatively high in authenticity as it may happen in real life that a person does not hear clearly or understand completely the full sentences spoken by his/her interlocutor and is hence forced to ask for clarification, repetition of information, and the like. Another example of an authentic task is Example 18, since it is very likely that a student would be asked those questions in real life, either in oral or written communication.

The following example, however, illustrates a task that is low in authenticity.

[Example 22] (Grade 6) **Write the past tense.**

Build _____
Catch _____
Go _____
Cook _____
Cut _____
Work _____
Try _____
Stop _____
Feel _____
Watch _____

The task in Example 22 is low in authenticity as students would most probably never find themselves in a situation in which they would be asked/required to provide isolated past tense forms of verbs in real-life communication. The

task instruction is further troubled by its failure to specify what past tense the verbs should be in and different answers are likely to be provided owing to the existence of different past tenses in English.

In determining whether a task is authentic, a number of factors should be considered, such as "the characteristics of test takers, of the TLU task, and of the test task" (Bachman & Palmer, 2004, p. 29). Moreover, according to Bachman and Palmer (2004, p. 28), given that authenticity is relative, we should speak of low or high authenticity of a task, rather than a task being authentic or inauthentic. Authenticity can be achieved and/or improved in several ways: by setting a task in such a way that it simulates a real-life situation or communication act, as illustrated by Examples 18 and 21; by contextualizing/personalizing instructions, i.e., by writing them in a way that invites students to imagine themselves in the given situation(s) or to observe someone being in the depicted context(s) (e.g., the instruction for the task in Example 18 could read: *Your friend is staying with an English family who are asking him the following questions. How might your friend respond to them?*; and by contextualizing/personalizing a task, i.e., by placing it within a meaningful context students can easily relate to and imagine themselves in (see Examples 24 and 54) (more on contextualizing/personalizing tasks and instructions is given in Section 4.4, point (6)).

Consider the following two examples. The first one, Example 23, is an example of a task that is low in authenticity as students are not invited to simulate solving a real-life task and they can hardly relate the task to their own selves, while the other example, Example 24, can be labelled as being high in authenticity since it situates the language task in a context that students can easily relate to and imagine taking part in. More information on authenticity and contextualization and more examples are provided in point (6) in Section 4.4.

[Example 23] (Grade 6) **Complete the sentences with the comparative form of the adjectives in brackets.**

Your cake is _____ (tall) than mine.
Who is a _____ (good) friend: John or Tom?
Are you _____ (old) than your brother?
Is your house _____ (big) than your neighbors'?
Sara is as _____ (beautiful) as Martha.
This time I got a _____ (bad) grade than last time.
Oh, the house is _____ (clean) than when I left. Have you cleaned it?

[Example 24] (Grade 6) **Your friend has just come back from a journey to Paris and Berlin. You want to find out what he/she thinks about the two cities. Complete the questions with the comparative form of the adjectives in brackets.**

Which city is _____ (clean) and _____ (green)?
Do you think Paris is _____ (easy) to move around than Berlin?
Paris is _____ (big) than Berlin, right?
Which city is _____ (good) to visit in summer when we're on holiday?
Is Berlin _____ (far) from Novi Sad than Paris?
Is the climate in Berlin _____ (bad) than the climate in Paris?
Is Paris really as _____ (beautiful) as everybody's saying?

As is obvious from the two example tasks above, both measure the same type of language knowledge (comparison of adjectives), but in different ways. While items in Example 23 are indeed contextualized, i.e., adjectives are not tested in isolation but within sentences, the sentences are not interrelated in the sense that they create a meaningful whole or story that students can relate to. On the other hand, in Example 24, the very instruction invites the test takers to personally take part in the given situation. All the items, i.e., sentences, in this task are interrelated and comprise a meaningful whole; what is more, the context within which the task is situated is easily relatable for the test takers themselves. In summation, it can be said that the adjectives are

tested in a context in both tasks, not in isolation. However, in Example 24 the entire task, not just its individual items, is contextualized and is, thus, more relatable for the test taker, making this task more authentic, as students may indeed encounter the depicted situation, or one similar to it, in reality.

The instruction can therefore impact the authenticity of its accompanying task. If it is clear, motivating, and inviting, and thus helpful in navigating the test taker through the task and enabling him/her to relate the task to his/her own life or real-life situations, it increases the authenticity of the task. In case the instruction provides insufficient information or when it does not invite the test taker to relate the given situation to real-life, not only will it potentially mislead the test taker, but it could easily fail to engage him/her in doing the task. Needless to say, the items in the task accompanied by an effective instruction need to be as tightly tied together as possible in order to elicit the language behavior students would exhibit in similar real-life situations.

4.2. Instructions and other test elements

Different authors agree that prior to creating a test, some planning needs to be done. The phases of planning, test design, and test administration are known as test development (Bachman & Palmer, 2004, p. 85). The same authors claim that the amount of time spent on test development depends on the situation; the more important the decisions to be made based on the results of a test are, the more time needs to be invested in its development. For that reason, many of the stages of the test development procedure may be skipped in designing certain teacher-made tests, while all of them need to be strictly adhered to in designing high-stakes tests, whose results will be used for making decisions exerting a major impact. However, Bachman and Palmer (2004, p. 85) warn that one thing should never be compromised regardless of how detailed the planning is — the qualities of the test: objectivity, reliability,

validity, authenticity, interactiveness, and impact (see Section 1.1. for more information on these).

Bachman and Palmer (2004, pp. 85–93) divide the test development process into three stages: design, operationalization, and administration. In the first stage, crucial elements of a test are defined and described, such as the purpose of the test, the description of the target language use domain, the characteristics of test takers, the definition of the construct, etc. The definitions and descriptions produced in this stage steer the process of test design, scoring, and interpreting test results. Needless to say, the test designer needs to ensure that the tasks included in the test are congruent with the definitions and descriptions produced at this initial stage of the test development. The second stage, operationalization, presupposes the design of the test, i.e., designing test tasks and writing instructions. The final, third, stage presupposes the administration of the test, the collection of information, and its analysis.

As can be seen, instructions are part of the test development procedure. Even more so, they need to correspond to the purpose and the construct(s) of the test and its specific tasks outlined in the initial stage of the test development process. In other words, there should not be a task requiring students to exhibit knowledge which was not planned in the design stage to be measured.

To illustrate the interrelatedness between the definitions and descriptions from the design stage and instructions, consider the following example task.

[Example 25] (Grade 6) **Describe your typical day.**

The first question that arises here is what the teacher wants to measure with this task: the knowledge of appropriate grammar structures (the use of the Present Simple Tense, adverbs of time, etc.), the knowledge of vocabulary for naming different activities, proper spelling, sentence construction (i.e., word

order), or something else? Without defining what the construct of the task is in the design stage, i.e., without specifying what is to be measured by this task, the teacher may find himself/herself applying different scoring criteria when grading different students' answers, which undoubtedly decreases objectivity as a critical test quality. It is of utmost importance that the same elements be measured and graded in all students' responses. In other words, the teacher should focus only on those elements that he/she has stated he/she shall be measuring in the specific task. When the construct of the task is defined, the instruction for the task needs to be written in such a way that it elicits the behavior intended to be measured. More precisely, we need to communicate to test takers what exactly they need to do so that they exhibit the behavior we are interested in measuring. For that reason, the instruction in Example 25 should be more specific so as to give clear guidance to students what is expected of them. For instance, the instruction could read:

(1) *Describe the activities of your typical working day — when you get up, how you spend your morning, when you go to school, what you do after school, etc.;*

(2) *Describe your typical day on the weekend — when you get up, how you spend your morning, afternoon, evening, whether you have some typical activities you always do on the weekend, what your family do*;

(3) *Describe your typical day during the summer holiday — when you get up, how you spend your morning, afternoon, evening, what your family do*, etc.

Being specific and giving clear guidance to students on what they are expected to do also motivates them and helps them perform better. What is more, in a task like this one, additional information regarding the length of the expected answer would also contribute to increased objectivity in scoring (see Section 4.3, point (6) for more information on length).

To illustrate this point further, consider the following example.

[Example 26] (Grade 7) **Fill in the gaps with the most suitable word.**

Rainforests of South Africa are the natural _____ of the Ioris.
Ghosts sometimes appear in that _____ house.
The great white shark is maybe the deadliest _____ of the animal world.
A large cat with no tail that lives in forests is called a _____.
Large grey African animals with a big head and fat body which lives near water are called _____.

Again, when checking how appropriate an instruction is, or, how precisely it directs test takers, we need to consider the construct for the task we defined in the design stage. What do we want to measure with this particular task? As is, the instruction in Example 26 invites the test takers to supply any word they know to complete the sentences. If this is the case — that students are expected to provide any suitable lexeme to complete the sentences, then there are several problems with this task. First, some sentences check students' topical knowledge, not the knowledge of a foreign language, and it is against one of the postulates of foreign language testing: you can test only what you have taught (see Section 1.1). It is doubtful whether teachers of English as a foreign language would really have taught (or would relevantly teach) the particular elements presented and intended to be measured in Example 26. Second, a number of words that would be a good fit for each sentence are possible, so there is a danger that the variety of options may affect objectivity (as teachers might not accept all answers as possible and correct for a certain reason, e.g., they are not native speakers and may be unsure whether an answer provided is actually a good fit). Third, in asking students to recollect any word they know to complete the sentences, the question arises as to what particular aspects of vocabulary the teacher is seeking to measure with this task: the singular or plural form of nouns, spelling, the meaning of words, students' vocabulary span, or something else? To make it clear to himself/herself and to students, the teacher needs to pay due attention to defining and describing various test

elements first as this has a significant effect on both the test design and the interpretation of its results. It is very unlikely that the intention of the teacher in task Example 26 was to measure students' general knowledge of vocabulary. Rather, the intention was probably to check the vocabulary covered in one specific unit. However, that intention was not clearly communicated to the students through the instruction analyzed. To aid this, the instruction might read as follows: *Fill in the gaps with the most suitable word from Unit 8* (the number of the unit in which the tested vocabulary would have been covered). By specifying the source of information that is checked in this activity the intention of the task, to check specific vocabulary covered in a particular unit, would be specified and objectivity increased thereby (as the instruction implies that only the lexemes from a specified unit are expected answers).

Another example that clearly illustrates the importance of defining and describing the key elements in the design stage and how that impacts instruction writing is given in the task of Example 18, which is followed by an explanation of the impact the instruction has on the task.

4.3. Component parts of instructions

To ensure they are in compliance with the definitions and descriptions from the initial stage of test development (see Section 4.2), instructions should contain several component parts. All of them should enable test takers to understand what is required from them and under what conditions they are taking the test; at the same time, those component parts should also enhance the test's quality. Yet few authors of textbooks on foreign language assessment write about the component parts of instructions. Exceptionally, Bachman and Palmer (2004) do outline the following elements that effective instructions need to contain: test purpose, language abilities to be tested, parts of the test and their relative importance, procedures to be followed for all parts of the

test, and the scoring method. Additionally, some authors (e.g., McKey, 2008) also stress the importance of specifying the audience.

(1) Test purpose

The authors believe that the purpose of a test, or of its different parts if each part has a different purpose, needs to be clearly communicated to test takers as it provides justification for giving the test and it contributes to fairness (Bachman & Palmer, 2004, p. 185). The same authors contend that even though the purpose of teacher-made tests is, more often than not, obvious to students, "it is essential that students clearly understand the particular use of each test" (Bachman & Palmer, 2004, p. 185). The authors advise that in cases in which testing is a common procedure of the teaching process students be familiarized with the purpose of testing as part of the general information about the subject or course, while in instances when testing is not a frequent activity, the purpose of each test needs to be communicated to students upon administering it (Bachman & Palmer, 2004, p. 185). While the purpose of high-stakes tests is well known to students since they have chosen to take a certain test or are required to do so, such as taking an entrance examination at a college, taking a final exam, etc., the purpose of a classroom test may not be so clear or obvious to students and, hence, needs to be specified.

(2) Language abilities

In the same vein, specifying the abilities tested in different tasks is considered to enhance students' motivation and the accountability of test use (Bachman & Palmer, 2004, p. 186). This is true of both teacher-made and large-scale tests. The authors believe that in classroom tests a simple label denoting the ability tested in each task often suffices, such as in the following example:

[Example 27] (Grade 5) **READING. Read the text, then answer the questions. Use full sentences.**

Hello! My name is Libby Johnson. I'm twelve years old and my birthday is on the 14th November. I live with my mother, father and sister in Bournemouth, a town in the Southeast of England. It's a nice town by the sea. I'm a student at Highfields School. I'm good at science, geography and maths, but not so good at art and sport. I'm interested in cooking, reading and swimming.

Example 27 illustrates that the label 'reading' informs the test takers of the ability, skill, or type of knowledge measured by the given task. In large-scale tests commonly taken by students of different educational backgrounds, a simple label is often insufficient (Bachman & Palmer, 2004, p. 186) and the instruction for the same task as in Example 27 could read: *This task tests your ability to understand a text. Read it, then answer the questions in full sentences.*

Non-technical vocabulary should be used in order to ensure improved comprehension, especially in large-scale tests when students may differ with respect to their language backgrounds (Bachman & Palmer, 2004, p. 186). For instance, instead of labeling the task of Example 27 as 'reading comprehension' or stating 'This task tests your reading comprehension,' simple labels like 'reading' or 'ability to understand a text' are often opted for. More information on specifying the type of language knowledge or skill tested is given in Section 4.4, point (2).

(3) Parts of the test

If a test is composed of several parts, with each testing a certain aspect of language knowledge, test takers need to be informed of the parts the test

contains, the number of items in each task, and the time test takers have at their disposal for doing each part. This test format is typical of large-scale tests and less applicable to classroom tests, so these pieces of information are generally not considered obligatory for tests designed for classroom use. Some classroom tests may contain the information on time in their instructions, and, should that be the case, it is more a suggestion to test takers of how much time they are advised to spend on doing a particular task than a requirement.

(4) Procedures

The information on the procedures to be followed is also a crucial component of effective instructions. This information relates to the order in which tasks are supposed to be completed, as well as to where and how test takers are to record their answers. In tests that consist of a number of parts (typically large-scale tests), it is, more often than not, important in what order the tasks are done. The information on the prescribed order should be clearly given to test takers, either in writing or orally. In classroom tests, however, the order of doing the tasks is generally not important, hence this information is often not necessary and thus not given in instructions. Additionally, students need to know where and how they are expected to record their answers, both in large-scale and classroom tests, for which reason this information should always be given. It should be stated specifically how students are to indicate a selected response, e.g., by circling, underlining it, or where they are to supply answers to a recall type of task — on the line, in a column, etc. Failure to provide such information through an instruction may have consequences on grading and test results. In their paper dealing with the quality of instructions in teacher-made tests, Glušac and Milić (2021) observed that the information on the procedure (how to indicate answers, where to record them, etc.) is very rarely present in classroom tests; in the relevant study, it was found to be given more often to fifth-graders and eighth-graders and less to sixth-graders and seventh-grades. Consider the following example:

[Example 28] (Grade 8) **Complete the text with *the* where necessary.**

My uncle is a traveller. He lives in _____ Netherlands, but very year he goes to a different part of the world. He has already been to _____ Africa and _____ North America. He has written a book about his journey across _____ South America: he started in _____ Chile, traveled along _____ Pacific, to _____ Peru, then crossed _____ Andes and went down _____ Amazon in a small boat. […]

This instruction does not give all the necessary information to the test takers regarding how to complete the task — it does not specify what to do when the definite article is not needed: whether to the leave the gap empty or to use a certain sign to indicate that the article is not used in that particular place. Not specifying this piece of information in the procedure leads to the test's decreased objectivity as, when grading the answers, the teacher would not know whether a gap left unmarked is a sign that the student did not know whether the article should be used in that particular situation or whether it is an indication that the article was not used intentionally. Any interpretation of what an empty gap means by the teacher leads to decreased objectivity. For all these reasons, the test designer needs to know in advance that some instances do not require the use of the definite article and state explicitly how to indicate such an answer.

Consider another example that illustrates a failure to give precise information on the procedure and the consequences of such a practice.

[Example 29] (Grade 6) **Write the correct order.**

and/live/Exeter/Mel/in/Barney
reading/I/interesting/am/book/an
Harry/is/now/TV/watching?
does/what/mother/do/your?
crying/you/are/why?

Firstly, as can be immediately seen from the layout of this task, no particular space is provided for writing answers. Also, there are two likely potential ways to complete the task. The first is to simply assign a number to each word and write it above a corresponding word and then, next to each string of jumbled words, write the numbers in the order in which the words should be arranged to make a grammatically correct sentence, just as illustrated below.

1 2 3 4 5 6
and/live/Exeter/Mel/in/Barney 4 1 6 2 5 3

The other way is to simply rewrite the words next to the string of words in which they appear in the order in which they should be arranged to provide a grammatically correct sentence, such as in:

and/live/Exeter/Mel/in/Barney	Mel and Barney live in Exeter. / Barney and Mel live in Exeter. / In Exeter Mel and Barney live.

How exactly to do this task is not explicitly communicated to the test taker and the question remains as to whether any method would be more preferable or acceptable for the teacher and whether that would in any way impact grading. Needless to say, if students are expected to write or rewrite something as part of their answer, then adequate space also needs to be provided. In this case, Example 29, no space is provided, which further complicates the students' decision regarding how to record their answers. Additionally, the fact that a number of answers for the same string of words are possible further aggravates the situation in the sense that certain test qualities could be compromised, including objectivity and reliability.

(5) Scoring method

Bachman and Palmer (2004, p. 189) have asserted that making test takers familiar with the scoring system through instructions is also a necessity. The authors advise that the scoring information be given as a general instruction

if all the tasks the test includes are scored on the same principle, while if individual tasks are scored differently, then the information on the scoring procedure should be part of the individual task's instructions. The same authors also suggest that test designers inform test takers regarding how their answers should be written and how they will be scored (answers other than selected responses, e.g., one-word answers, or answers in the form of a sentence). More precisely, regarding cases in which written answers are expected, students need to know which aspects of their answer(s) will be measured and how long their answers should be. Consider the following example.

[Example 30] (Grade 5) **Answer the questions in full sentences.**

How old are you?
Are you tired?
What is your favourite colour?
Has your best friend got a pet?
Have you got a watch?

It would be worth specifying to students what in particular is intended to be measured in this task — grammar, vocabulary, sentence structure, or something else, as students would then surely devote more attention to the elements measured than to those that are not to be taken into consideration when grading.

In dealing with this specific constituent part of instructions, Bachman and Palmer (2004) do not make a difference between classroom and large-scale tests, from which it can be inferred that the same principle of making the scoring system known to the test taker should be applied to both types of tests.

To enhance the understanding of the instruction and what is required from students, the test designer might think of including an example (Bachman

& Palmer, 2004, pp. 183–184; Heaton, 1990, p. 170; Purpura, 2004, p. 128). Heaton (1990, p. 170) explains that examples should be provided only if the testing technique is not familiar to the test takers, which in the classroom situation is, hopefully, a rarity, as Brown (2000, p. 410) and Weir (2005, p. 54) warn, stressing that teachers should include in their tests only those tasks students are familiarized with. To illustrate the use of an example as part of the task instruction, consider the following examples.

[Example 31] (Grade 7) **Complete the sentences with question tags.**

You're studying English, aren't you?

John won the competition, _____?
I'm playing well at the moment, _____?
Your sister can't come to the party, _____?
It was raining yesterday, _____?
They'll help us, _____?

[Example 32] (Grade 7) **Read the sentences. Underline the correct preposition.**

He didn't listen to / for me.

She talked *to / at* her friend about the problem.
He always works *by / with* talented actors.
She likes thinking *of / for* difficult mathematical calculations.
They came *from / with* a poor region of Italy.
He doesn't worry *for / about* the dangers of his job. […]

As can be seen from Examples 31 and 32, in both tasks an example of how to do the activity is given. Additionally, Example 32 not only illustrates what students should do (choose the correct preposition), but how to do it as well (by underlining). Giving both types of examples can be worthy in case the test is taken by a student who has not attended the class aimed at preparing students for the test and is thus not familiar with certain types of tasks and

instructions prior to test administration. Also, giving an example illustrating how the task is supposed to be done improves the comprehension of the instruction as well.

However, it cannot go unnoticed that the instruction in Example 31 is flawed in two ways, as will be discussed in Section 4.4, points (2) and (4). First, it contains technical language — 'question tag' (as does the instruction for the task in Example 32), which the students might not be familiar with, for which reason they may provide different answers checking the validity of the statement coming before the question, including *Right?*, *Correct?*, *Yes?*, and the like, which is not the intended purpose of this task. Second, the instruction suffers from not being informative enough, which could jeopardize certain test qualities. It should remind the test taker that a question in an appropriate tense is needed, which could probably serve as good enough guidance to prevent the test takers from providing answers like the ones mentioned above.

(6) Audience

Besides these components, the existing literature on test construction mentions one more element that effective instructions for writing tasks should include: the audience. McKey (2008, p. 251) asserts that instructions need to specify the audience, the purpose of the task, and the scoring method. To this, Weigle (2009, p. 103) adds that instructions for writing tasks should also include a specification of length. Information on the audience for whom students are writing impacts the choice of their vocabulary, register, and grammatical structures, as well as the content. Moreover, by specifying the audience, the test designer contextualizes the task, which elicits the exact language the test designer intends to measure. To illustrate this point, consider the following example.

[Example 33] (Grade 7) **Complete.**

List three things you do that make you happy:
List three things that cause you stress (make you feel stressed out):
List three things parents can do to help their children stay healthy:

As can be seen, the instruction, among other components parts, also lacks the specification of audience. If this information was offered, students would know who would be, in theory, reading their answers and would likely do their best to write appropriately for that particular audience. In attempting to complete the task illustrated in Example 33 and in listing the required things, students could have different audiences in mind, such as friends, teachers, etc., as well as different domains of life, such as school, life in general, traveling, and the like. When specifying the audience and the context, the test designer increases both the authenticity of the task and the chances of eliciting the language he/she intends to measure, i.e., specific vocabulary, register, grammatical structures, and the like. When the task instruction does not contain adequate information regarding the audience and context, it is then open to students' interpretation and the number of possible answers increases, some of them certainly representing answers the test designer had no intention of measuring. Moreover, such an instruction fails to help test takers relate the task to their own life or a real-life situation.

In considering what information should be included in test instructions, Bachman and Palmer (2004, p. 190) claim that test designers should be guided by two basic factors: (1) how familiar the test takers are with the tasks, and (2) the number and variety of tasks the test includes. As mentioned earlier in this section, in a classroom testing situation some information enlisted in the given review of literature can be omitted from test task instructions if it has already been communicated and is well known to the students, such as the purpose, examples, the scoring method (in cases in which the same method applies to all the tasks), etc. Conversely, all this information needs to be given

in large-scale tests since students of different backgrounds take them and may not be familiarized with these particular elements.

4.4. Features of instructions

A comprehensive review by Glušac and Milić (2021) of the literature the students receiving instruction in foreign language assessment are commonly referred to led them to single out a number of features effective test task instructions should possess. They relate to the following aspects: (1) length, (2) language, (3) form, (4) informativeness, (5) inclusion of component parts, (6) additional features, and (7) visibility.

(1) Length: Task instructions should be short in length.

Many foreign language assessment textbook writers (e.g., Bachman & Palmer, 2004; Dimitrijević, 1999; Purpura, 2004; Weir, 2005) emphasize that instructions should be short and that only essential information should be given, in as precise terms as possible. As suggested by Bachman and Palmer (2004, p. 190), reading instructions should not take up too much time which students could otherwise spend on doing the test. Consider the following example of a short instruction.

[Example 34] (Grade 5) **Circle the correct answer:**

Summer is *the warmest/warmer* season.
An elephant is *smaller/the smallest* than a mouse.

The instruction for this task is short in length and gives sufficient information regarding what is expected of students.

Sometimes, though, instructions are inappropriately short, thus lacking crucial information. To illustrate this point, consider the following example.

[Example 35] (Grade 5) **Complete the sentences.**

	Ben	**Katie**	**Ravi**
go climbing on Tuesdays	F	T	F
watch TV on Saturdays	T	F	T
play football	T	F	F
go rollerblading in the park	T	T	F

Katie _____ climbing on Tuesdays.
Ravi _____ TV on Saturdays.
Ravi _____ rollerblading in the park.
Katie and Ravi _____ football.
Katie _____ TV on Saturdays.
Ben and Katie _____ rollerblading in the park.

When the task given in Example 35 is analyzed, it can be concluded that its instruction is short, yet inadequately so, since test takers are not instructed with respect to all the steps they need to carry out in order to do the task, such as *look at the table* or *study the table*, as well as regarding what particular element(s) of knowledge is needed for the completion of the sentences, etc. When instructions are insufficiently short, they can be labelled as being insufficiently detailed, a feature more extensively dealt with in point (4) in this section.

(2) Language: Task instructions should be written in simple, non-technical, unambiguous, clear, and intelligible language, which can be either the students' mother tongue or the relevant foreign language, or sometimes even a combination of the two languages.

A number of authors (Bachman & Palmer, 2004; Heaton, 1990) advise that instructions should be written in simple, non-technical language so that all students can understand them easily. For instance, technical words such as *verbs*, *nouns*, *adjectives*, etc., should be replaced with *words like the following* (followed by example words) (Heaton, 1990, p. 169). The same is true of

more technical words such as *comprehension*, which may be replaced with *understanding* (Bachman & Palmer, 2004, p. 186). Consider the following example that illustrates this point.

[Example 36] (Grade 6) **Complete the collocations.**

_____ dinner	_____ your teeth
_____ to bed	_____ a photograph
_____ an e-mail	_____ the shopping
_____ the ball	_____ your pyjamas

There are a number of problems with this instruction, but first and foremost is the problem related to the issue being discussed in this section — technical vocabulary. Namely, the instruction contains the word 'collocations,' which is a rather technical term, and it is questionable as to whether sixth-graders would really understand what it means. It should be replaced with a more understandable word or phrase, as in: *Complete the gaps below with words that typically go with the words given, such as make, do, take, etc*. Another problem with this instruction is that it does not specify which particular answers are expected. Hence, the teacher would potentially receive answers he/she had not covered with the students. This, as explained earlier in Section 4.1, can affect the test's qualities. If objectivity is to be ensured, a good key, including all possible answers, should be made, or the students need to be referred to a specific unit where the words that are intended to be measured by this task have been covered.

Given that students in a single class can vary greatly with respect to their foreign language proficiency, technical terms should not inherently be excluded at all times and for all students. Moreover, if a test measures different elements of knowledge and a number of language abilities, it could be appropriate or necessary to specify what particular element or ability is tested by a certain task or a cluster of tasks by stating its name. Purpura (2004, p. 127) also suggests that an indication of the grammatical ability being tested should be

made part of a test task instruction. Consider the following example as an illustration of how such information can be included in the instruction.

[Example 37] (Grade 5) **GRAMMAR. Complete the sentences with the right form of BE.**

Where _____ the Thames and the Mississippi?
A: Are you OK?
B: Yes, I _____.
Where _____ your school?
Who _____ Mark and Joe? […]

As can be seen, the instruction is devoid of any technical terms, but is preceded with the label 'grammar,' informing students what particular element of knowledge or skill is tested in this very activity. Such a practice can help students in planning their execution of the test, since, when tasks are clustered and labelled according to the area of knowledge or skill they measure, students can decide what part of the test to take in what order. Also, such a label might be an extra piece of information on what specific knowledge they need to recall to do the activity, as the line between grammar and vocabulary, for instance, can sometimes be rather subtle.

Another example of how the ability or type of knowledge that is tested can be formulated is as follows:

[Example 38] (Grade 5) **Present Simple: yes/no questions**
Match the questions to the answers:

Do you live in Russia?	Yes, they do.
Does Mikey speak French?	No, we don't.
Do elephants live in the Arctic?	Yes, she does.
Does Jessica play a musical instrument?	No, they don't.
Do people speak English in Britain?	Yes, he does.
Do we have homework for tomorrow?	No, I don't.

Example 38 illustrates an instruction that is devoid of technical terms, but which does contain a statement of the type of knowledge tested in this particular task (*Present Simple: yes/no questions*) and, hence, still informs students about what area of knowledge is tested, which they could find relevant.

The failure to inform test takers about the type of language knowledge or skill measured by a particular task can be misleading, in the sense that students are not properly informed about how to formulate their answers, as in the following example.

[Example 39] (Grade 6) **Write down:**

A request:
A thing you are not allowed to do:
A thing you can do:

As can be seen from this example, a simple label denoting what is measured by this task would be helpful as students would know what particular knowledge or ability is intended to be tested — e.g., whether it is sentence construction or vocabulary. Successful completion of this task could seemingly be achieved by simply providing individual words (e.g., *ski*, *yell*, *be late*, etc.) or full sentences (e.g., *I'm not allowed to go skiing, to yell, or be late*). Without knowing what particular knowledge or skill is intended to be measured here, the test taker might not know how to formulate his/her answers. Listing the ability or knowledge element in this case would also communicate the purpose of the task to test takers. When all this is taken into consideration, the instruction could read as follows:

GRAMMAR. Write an example sentence for the following:

A request:
A thing you are allowed to do:
A thing you can do:

In this way, students would be informed that it is their grammar being checked in this activity. In other words, the teacher might be interested in assessing students' knowledge of modal verbs used for the given purposes (request, permission, ability), or other sentence structures used for expressing the same purposes. The added label indicating the type of knowledge or skill measured would presumably help the test takers decide on the form of their answers. The necessity to include the information regarding the type of knowledge or skill tested was previously discussed in Section 4.3, point (2).

In the same vein, instructions need to be written in unambiguous, clear, and intelligible language since they need to give clear guidance to students regarding what they are expected to do. Failure to communicate the intention is clearly likely to result in students' poor performance and the acquisition of inaccurate test results. To illustrate this, consider the following example.

[Example 40] (Grade 5) **Change the nouns into the plural using the word in the brackets.**

I had a dog. (four) _____
Larry sold an old table. (several) _____
David will buy a new car. (three) _____
My friend bought a cake. (two) _____

The instruction for this task is ambiguous and not very clear as it does not instruct the students precisely in what they need to do. To illustrate, as it is worded, the instruction may raise a number of questions for the students, such as: Which nouns do I need to change, for instance in *My friend bought a cake* (*friend* or *cake*), etc.?, Do I need to replace the nouns with the words in brackets and make the words in brackets plural?, Do the words in brackets help me in any way change the nouns from the sentences into plural?, etc. For all these reasons, we can assume that some students would find the instruction not clear or intelligible enough. Additionally, the instruction is insufficiently

detailed (see point (4) in this section) as it does not inform the students that they need to rewrite entire sentences with the changes required. If students were to write on the line only what is requested in the instruction — the use of the plural form of the noun accompanied with the determiner in brackets, such an answer would not reveal whether the students know that the determiners in the original sentences do not co-occur with the ones in brackets and should be left out when the noun is in plural. The same instruction is also flawed in two other ways. First, it is given in the form of a complex sentence (see point (3) in this section), for which reason it could be too difficult, unclear, or unintelligible for fifth-graders to understand, and it presupposes their recognizing nouns before doing the very task. Dimitrijević (1999, p. 95) warns against the practice of designing such tasks in which responding to a certain question/task is dependent on responding to the one preceding it. Here, this means that if students were to fail to identify a noun first, they would not be able to do what the instruction requires them to do. That further raises the question of the purpose of the task: whether it is to recognize nouns, to exhibit one's knowledge of plural forms of nouns, or both. If the purpose was to check only the students' knowledge of plural forms of nouns, the task could be improved by underlining the nouns that the teacher wants the test takers to use in the plural form.

Another example of an instruction that is ambiguous, or not clear or intelligible, is the following:

[Example 41] (Grade 5) **Rewrite the complete sentence using the adverb in brackets in its usual position.**

He listens to the radio. (often)
They read a book. (sometimes)
Pete gets angry. (never)
Tom is very friendly. (usually)
I take sugar in my coffee. (sometimes)

The instruction is problematic as it is unclear what the 'usual' position of the given adverbs is, since the position depends on the meaning we want to convey. In all likelihood, fifth-grade students are probably not aware of the different pragmatic values different positions of adverbs can have, so the test designer should certainly not punish them if they place the adverb in a position that does not match what the designer labels as 'usual,' but which is indeed an acceptable usage/position of the adverb in question. Additionally, the instruction in Example 41 contains technical language — 'adverb.'

Another language-related issue that different authors (e.g., Bachman & Palmer, 2004; Heaton, 1990) have addressed is that of which language is appropriate for giving instructions: students' mother tongue, the relevant foreign language, or a combination of the two. Weir (2005, p. 57) and Heaton (1990, p. 169) agree that in case a complex instruction is needed to explain what test takers are supposed to do, the most viable solution is to give it in the test takers' native language. Weir (2005) does acknowledge, though, that such a practice may be criticized by many, while Heaton (1990) asserts that such a practice should be employed only at the elementary level of knowledge. Along the same lines, Bachman and Palmer (2004, p. 182) also suggest the students' native language could be used in writing instructions if the test designer believes any kind of misunderstanding could arise if instructions were to be given in the target language. Considering all this, it can be concluded that test designers should not fundamentally try to avoid the use of the students' mother tongue for writing instructions; rather, they should also take into consideration the level of complexity of the instruction, as well as the difficulty of the language used in writing it. The practice of giving an instruction in both languages is also not uncommon (Glušac & Milić, 2021), and it might be a way to make the task and test more adequately approachable to students with diverse language proficiency levels. Consider the following example with the instruction in two languages.

[Example 42] (Grade 5) **Write the words in the right order. (Reči nisu po redu. Napiši reči po redu da dobiješ tačnu rečenicu.).**

long/is/How/the river Danube? _____
is/big/How/Belgrade? _____
high/how/is/Mount Everest? _____

(3) Form: Task instructions should be written in the form of simple sentences.

In situations when the task requires a complex instruction, Heaton (1990, p. 169) advises that the instruction should be written in the form of a few simple sentences, each guiding students regarding how to perform a single step of the entire activity.

Consider the following example illustrating an instruction given in the form of a complex sentence.

[Example 43] (Grade 8) **Use the given phrases to complete the sentences making any necessary changes to the given phrase.**

take place take your time take part in take the exam

You don't have to hurry. _____!
I studied hard. I will _____ as soon as possible.
The party _____ next Saturday at 10 pm.
Peter and Helen _____ in the school play last year.

Even thought it might seem that the instruction is short, as it is given in the form of a single sentence, the sentence is complex, containing a non-finite clause (*making any necessary changes to the given phrase*), which might be too difficult for students to understand. In this case, a better option would be to split this sentence into two simple sentences, each informing students about a single action they need to perform. For instance, the instruction could

read: *Use the given phrases to complete the sentences. Make any necessary changes to the given phrases so that they fit into the sentence correctly.*

(4) Informativeness: Task instructions should be informative, i.e., sufficiently detailed.

A number of authors (e.g., Bachman & Palmer, 2004; Cohen, 1994; Weigle, 2009; Weir, 2005) point out that instructions should be sufficiently informative, in the sense that they need to tell students exactly what they are expected to do. Failure to provide test takers with complete information regarding what they should do will likely force them to spend valuable time on studying the task and inferring this on their own. Furthermore, in order to avoid coming to such a conclusion on their own, which could result in misinterpretation, students will often disrupt other students by asking either the teacher or their classmates for clarification. Consider the following example.

[Example 44] (Grade 5) **as … as**

skiing/interesting/athletics
London/beautiful/Glasgow
Motorbike/fast/cars
February/cold/December
flying/dangerous/driving

As can be seen in Example 44, the instruction is not informative enough, as it specifies only the element of grammatical knowledge students are expected to show mastery of, but it lacks any information regarding what students should actually do with it. For that reason, the test takers would probably spend some time only in figuring out for themselves what they should do, or, alternatively they might ask either the teacher or a classmate for clarification, and in doing so likely disrupt others. It must be further emphasized that in students' reaching of a conclusion on their own, the question remains as to

whether it is the right one, i.e., whether they have figured out correctly what the task requires them to do, owing to which they could underperform.

Another example that illustrates an insufficiently detailed instruction is the following:

[Example 45] (Grade 6) **Complete the sentences.**

Kopaonik is a _____. Kilimanjaro and the Alps are _____, too.
Canada is a _____. Brazil and Mexico are _____, too.
The Mississippi and the Thames are _____. The Nile is a _____, too. […]

The instruction is insufficient in that it does not inform the test takers what specific element of knowledge is elicited for completing the sentences. Without this being specified in particular, students would likely spend a certain amount of time on trying to figure out what is needed from them. Moreover, there are numerous possibilities to complete the given sentences, such as *Kopaonik is a great mountain. Kilimanjaro and the Alps are exceptional, too*, which further aggravates the situation.

One more illustration of an insufficiently detailed instruction is the following.

[Example 46] (Grade 5) **Correct the mistakes in the following sentences.**

Last night, Samantha have pizza for dinner.
My pet lizard was died last month.
Yesterday I spend two hours cleaning my living room.
This morning before coming to class, Jack eats two bowls of cereal.
What was happened to your leg?

This example shows an instruction that fails to tell the students what mistakes to look for, i.e., what area of language knowledge the mistakes in these sentences

fall into. Given that the instruction is not accompanied by the specification of the type of knowledge tested, the test takers would spend valuable time on deciphering where the mistake is in each sentence and how many mistakes there are in each instance. They could also disrupt the test administrator by asking questions for clarification, or they might provide answers other than the ones sought since the instruction is not sufficiently detailed.

(5) Inclusion of component parts: Task instructions should contain all the necessary component parts or pieces of information, such as scoring, examples, time, purpose, audience, language/abilities listed, procedure, etc., just as discussed in Section 4.3.

In addition to documenting test task instructions lacking a certain element, the study of test task instructions conducted by Glušac and Milić (2021) revealed that it is not an uncommon practice for a task instruction to be missing entirely in teacher-made tests. Instead of the instruction, at times only the ability/type of knowledge intended to be measured was given. Similarly, Fleming and Chambers (1983, cited in Marso & Piggie, 1991) reported that instructions were entirely missing from a third of the tasks they analyzed and their corpus, which included 342 tests for grades one through twelve. Consider the following examples:

[Example 47] (Grade 8) **First conditional.**

If we _____ (go) to London, we _____ (visit) the Tower.
If the shop _____ (be) open, I _____ (buy) you a souvenir.
If she _____ (miss) the bus, she _____ (get/not) here on time.
He _____ (help) me if he _____ (know) the answer.
Jane _____ (let) you in if I _____ (be/not) here.

[Example 48] (Grade 5) **Places in town!**

for a bus? …….
for a drink? …….
for a swim? …….
for books? …….
to see a film? ……..
for a walk? …….

As can be seen in the examples above, only the type of language knowledge that is tested is labelled, while guidance regarding what exactly is to be done is completely missing. Failure to instruct test takers as to what exactly they should do can have adverse consequences on test performance. For instance, in Example 47 the type of knowledge listed might not be informative enough for students, especially as it is given as a technical term. In Example 48, on the other hand, students are not referred to any specific unit in which the vocabulary tested by this activity was covered, if vocabulary is at all the intentioned object of measure of the task, since this is also not clear. Failure to give students such particulars can lead to their providing a myriad of answers the teacher might have had no intention of measuring. Additionally, it compromises objectivity and reliability.

Consider the following example as an illustration of a task which lacks instruction entirely:

[Example 49] (Grade 6)

She stayed in space for three days.
? _____
- _____

She completed some experiments.
? _____
- _____

It landed in Kazakhstan.
? _____
- _____

A task not accompanied by an appropriate instruction requires students to spend valuable time attempting to figure out what they need to do and how. Depending on the task, this could result in them arriving at a conclusion regarding what to do that is not in accordance with the teacher's intended purpose for the particular ask. Hence, students would likely underperform and then possibly be punished by not earning many or any points in that task for something that was no fault of their own, but of the teacher who designed and evaluated the test.

(6) Additional features: Task instructions should be language-accurate, correct, and personalized/contextualized.

In addition to the features (1)–(5) singled out in their review of assessment-related literature, Glušac and Milić (2021), in analyzing the corpus that included teacher-made tests for the purposes of their study, also identified three additional features not addressed sufficiently or at all in the consulted literature: language accuracy, correctness, and contextualization/personalization. The authors do, however, acknowledge that further investigation is needed as to how these features might affect test performance.

The language of instructions needs to be accurate in the sense that instructions should be free from any language issues. In their analysis of 308 instructions for grades five through eight of the primary school, Glušac and Milić (2021) discovered a number of language issues, including misspelt words, incorrectly used articles, lack of punctuation, the lack of diacritic symbols in instructions in Serbian, etc. This finding corroborates that of Fleming and Chambers (1983, cited in Marso & Piggie, 1991), who, in their analysis of 342 tests,

found that 15–20% of the analyzed instructions contained a language issue (grammatical, punctuation, or spelling error). Instructions written in poor, erroneous language could affect students' performance and cause further issues by serving as a faulty source of linguistic information that students might utilize (Glušac & Milić, 2021). Consider the following example as an illustration:

[Example 50] (Grade 6) **Match (povezi):**

My	well
I'm	ear hurts
I don't feel	sick
I feel	a headache
I've got	hungry

In Example 50 there are obvious language issues stemming from the Serbian translation of the English instruction given in brackets. More precisely, there are two language issues: the lack of a diacritic symbol and capitalization. The one-word instruction should be written with a diacritic symbol accompanying the letter *z*, i.e., it should be *ž*. In the Serbian language, the use of letters without their accompanying diacritic symbols is considered to be erroneous; hence, its use in a language test could indicate to test takers that the omission of those symbols is acceptable and can motivate their own use of it. Moreover, the instruction in Serbian should begin with a capital *p*, to mark the imperative sentence beginning, as well as to unify it with the instruction in English, which is capitalized. Additionally, there is one more language issue here, not in relation to the instruction, but the task itself. Namely, the second half of each sentence lacks a full stop, thus reflecting an error in punctuation. Using language that is contrary to the norms of a particular language should be avoided as it could function as an erroneous source of linguistic information to test takers, whose errors they may then go on to perpetuate.

Additionally, Glušac and Milić (2021) discovered a number of instructions in their corpus that were incorrect, i.e., they gave misleading information to students. Consider the following example:

[Example 51] (Grade 5) **Order the questions.**

there/is/a/in/television/the kitchen
you/three/got/sisters/have
she/playing/tennis/does/like

The instruction in Example 51 is misleading since the test takers are not expected to order the questions in, let's say, the order of preference, but to order the words in each string so as to form a question. Most students are probably familiar with this testing technique and are not likely to have a problem when doing this task; still, if this is a test testing students' language ability, then it should itself contain exemplary language — correct, unambiguous, and appropriate. Otherwise, teaching and testing would not be congruent. While language teachers generally strive to facilitate students' language accuracy, on the one hand, they as test designers sometimes showcase contradictory language carelessness on the other hand, something that they must be committed to avoiding.

Another example illustrating a misleading instruction is the following:

[Example 52] (Grade 6) **Rewrite the sentences using the correct word.**

I am/are/be/is from the USA.
Anne and Tom am/are/be/is feeding the cows.
Joe drive/drives a school bus.
Our friend sometimes go/goes to the supermarket.
Andy am/are/be/is washing his pullover. […]

The instruction is misleading in that, according to the format of the task (no lines for writing answers are provided nor is there enough space for recording

answers), it can be concluded that the students should choose and indicate a correct answer, not really rewrite an entire sentence with the chosen verb form as is instructed. Giving a misleading instruction in this task could compel students to spend valuable time on really rewriting the sentences with the verb form they consider to be the most appropriate. In similar situations when a misleading instruction is given, students can underperform or give answers the teacher has to some degree caused, but which he/she had not expected them to give, for which reason the objectivity, validity, and reliability of the entire test are jeopardized.

In their extensive research on the quality of instructions in teacher-made tests, Glušac and Milić (2021) discovered that a certain number and type of instructions would, in a way, set the scene or describe the context in which the activity in the particular task happens. Additionally, all the items in such a task are centered around a particular event described in the instruction. The authors came to the conclusion that such an instruction and the accompanying task might boost students' motivation as such a task would seem to be more engaging than the one whose common instruction in the imperative form does not tie together the task items into a meaningful contextual whole (referred to as *contextualization*) or relate them to the test taker himself/herself (referred to as *personalization*). Personalized instructions and tasks are in fact contextualized instructions/tasks that relate to the test taker himself/herself or someone close to him/her (i.e., a friend, a parent, a sibling, etc.), rather than to some other people the test taker does not know personally. In a task with a personalized instruction, the test taker is invited to act as he/she would in a similar real-life situation or to imagine a person close to him/her taking part in it (see Example 24 and Example 54 as illustrations of personalized tasks). As discussed earlier in Section 4.1, personalized/contextualized tasks increase the task's authenticity and thus enable the test designer to generalize beyond students' test performance (Bachman & Palmer, 2004, p. 24). Glušac and Milić (2021), however, identified an alarmingly small number of such

contextualized or personalized instructions in their corpus. The instructions contained in the analyzed corpus of tests, in contrast, too often corresponded to tasks with unrelated items, seemed overly formal, appeared to be written in a hurry, or failed to establish any sort of context relevant for the test taker. Consider the following example:

[Example 53] (Grade 7) **Put *should/shouldn't.***

You _____ tell jokes.
You _____ help your friends.
You _____ cheat in tests.
You _____ injure other people.
You _____ drink alcohol.

Even though the instruction in Example 53 is written in accordance with many of the features previously discussed (it is short, simple, clear, unambiguous, is not misleading, does not have language issues, etc.), it is not very motivating or engaging. Bachman and Palmer (2004) state that "[t]est instructions also serve as an important affective goal: motivating students to do their best" (p. 182). Therefore, the students would probably be more motivated by the instruction and task in Example 53 if they could relate the knowledge measured by that activity to a real-life situation or their own life. One way to increase the test takers' engagement and interest in the task is to contextualize or personalize it, which would almost surely not diminish the quality of the items or the importance of the activity, but rather only increase the test takers' motivation. Contextualizing or personalizing the instruction and the task increases the task's authenticity, as already discussed in Section 4.1. When contextualizing/personalizing a task, the context in which items are placed needs to be clear and unambiguous in order to enable the student perform at his/her best. As an illustration, the first item in Example 53 is not contextualized properly, as there are situations in which it is more or less desirable to tell jokes, so students may find the given context imprecise or ambiguous.

Consider the following example illustrating how the same element of knowledge (modal verb *should* and its negative form) intended for assessment in Example 53 could be presented for the purpose of assessment in a more interesting, personalized, and authentic way.

[Example 54] **Your friend who lives abroad is visiting Serbia for the first time. Give the friend some advice by completing the sentences with *should* or *shouldn't.***

You _____ try some of our national dishes, such as barbecue.
You _____ visit both Belgrade and Novi Sad.
You _____ rent a car, as the public transportation is really good.
You _____ have Serbian dinars with you because we don't accept euros.
You _____ stay in a hotel because there are other good places that are much cheaper.

As can be seen in Example 54, contextualization has been achieved by personalizing the instruction and writing items that relate to the test taker himself/herself. The instruction is also contextualized in the sense that it sets the scene for the story, event, or action all the items relate to; both the instruction and the items are, thus, tied together, creating a story that bears relevance for the test taker. The suggested instruction is considerably longer than the original instruction in Example 53, but it contains simple language that a seventh-grader would likely understand with ease and it does not take too much time to read. It does not contain any technical language, is sufficiently informative, and is almost certainly more engaging than the one in Example 53. A test designer might contextualize/personalize a task which lends itself to being placed in a meaningful context with the intent of improving students' engagement and motivation. Such contextualization/personalization of a task must also be aligned with the intended purpose of the task. As Bachman and Palmer note (2004), "[c]ertain test tasks may be

relatively useful for the intended purposes, even though they are low in either authenticity or interactiveness" (p. 29).

Relevant contemporary literature mentions two types of contextualization: item (e.g., Douglas, 2010) and task (e.g., Bachman, 1995; Weir, 2005) contextualization. The former implies placing the tested elements of a language within a meaningful context (e.g., a sentence — see Example 53) and not testing them in isolation as if they do not have any relevance to real language use (see Example 4 or Example 5 in Chapter 3). Other than that, when contextualizing/personalizing, a teacher/test-designer could create a task that closely resembles a situation which would be easily encountered in actuality in order to measure how successfully test takers can apply their language knowledge or skills to manage the given situation by prompting them to do so. By doing this, the students are compelled to imagine themselves or somebody they know well being in the described situation (in case the task is personalized) or to observe somebody else not known to the test taker being in it (in case the task is contextualized), and then complete the task as if participating in the given situation or watching it unfold before their eyes (see Example 54). A task that requires the test taker to respond in a way in which he/she would act and use the foreign language knowledge in a real-life situation can be high in authenticity (see Section 4.1). In summation, contextualization can be achieved by testing each language item as part of a larger unit of text (e.g., expression, sentence), not in isolation, and through tasks that call for students' language use in the sense of their own creation of an answer, rather than simply calling on them to provide those he/she has learned verbatim. Dimitrijević (1999, p. 59) claims that the more a test is contextualized, the worthier and more reliable the results we get.

To the book's author's best knowledge, in relevant contemporary literature there is virtually no mention of the concept or practice of the contextualization/personalization of task instructions, regarding the third possible type of

contextualization mentioned and exemplified in Example 54. There is some indication (cf. Glušac, Milić, & Lužajić, 2021) that such test task instruction contextualization would be beneficial — instructions that not only inform test-takers what is expected of them, but also set a scene, i.e., describe a situation in which students are invited to take part either as participants or observers of others taking part in it (see Example 24 and Example 54). The items that follow such an instruction, if constructed appropriately, represent a meaningful whole in the sense that they are all relevant to the situation described in the instruction, while the task also serves as a simulation of a real-life event the test taker is invited to take part in. There is some evidence that such instructions may enhance test takers' motivation for taking the test and their performance (cf. Glušac et al., 2021), but the notion and practice of instruction contextualization/personalization should be further explored in order to obtain more conclusive evidence towards gaining a clearer and more substantial understanding of the true effectiveness of such instructions.

(7) Visibility: Instructions need to be clearly distinguishable and visible to students.

In making instructions clearly visible and distinguishable, perhaps the best option is to bold them, so that they stand out in comparison to the other text in the test, though other visual presentations like underlining, italicizing, or manipulating the font size could also be effective. Alongside such distinctive markers, space also needs to be planned and provided for writing answers, such as lines, boxes, etc. If students are expected to circle or underline an answer, then appropriate spacing should be created between the items in such a task. When appropriate and possible, interesting typography should be used and the tasks could be accompanied and illustrated with a few related pictures, so that the test does not seem too formal, rigid, uninteresting, and unrelated to real life. All in all, it is not only the content of the test that matters and that needs to adhere to certain rules and principles, but the layout of the

test as well, and its tasks should also be clear, well-organized, and appealing. By constructing the test this way, students might feel less anxious and more motivated to do the tasks. Consider the following examples as an illustration of how the test layout may affect some aspects of test performance.

[Example 55] (Grade 8) **Complete the descriptions.**

Jack is a re student. He always does his homework.
The new teacher's very fr . She smiles a lot.
Angela isn't fat. She just isn't very sl . […]

When analyzing the instruction for the task in Example 55, it appears to be insufficiently informative, as it does not really provide adequately detailed information to students — which descriptions, where they need to be completed, etc. When the test taker would attempt to analyze the task to infer further information, not much would likely be revealed to him/her, as the layout of the task is unclear. Namely, the student would not know where exactly the sentence/description is incomplete. There is no line, for instance, indicating the place where the expected answer is to be provided. The student might think that the space between the two letters and the full stop in the first item is a typing or printing error, not the place where the answer should be recorded.

Another example that can illustrate how the layout might impact test performance is the following:

[Example 56] (Grade 6) **Correct the mistakes in the following sentences.**

Last night, Samantha have pizza for dinner.
My pet lizard was died last month.
Yesterday I spend two hours cleaning my living room.
This morning before coming to class, Jack eats two bowls of cereal.
What was happened to your leg?

This example task does not specify how the mistakes should be corrected (by underlining them and writing the correct form above the mistake or after the sentence, by rewriting the entire sentence, etc.), while the layout does not provide any help, i.e., it does not offer any clear clues, to students that would allow them to infer how to record their answers. Namely, there are no lines provided that might suggest that entire sentences need to be rewritten with the mistakes corrected, nor is there appropriate spacing between the sentences so that the students could write the correction above the problematic spots. As mentioned before, failure to format a test and its tasks accordingly can lead to confusion, the asking of clarification questions and thus the disrupting of other students, underperformance, etc.

In teacher-made tests, the silver lining is that imperfections that a test might suffer from are repairable on the spot, for which reason students are typically not significantly affected by any adverse consequences of such tests. In standardized testing, however, failure to provide precise instructions, an error in the task or key design, and other issues are not recoverable and generally affect test takers in a negative way. In standardized tests, invigilators are not allowed to provide any clues, additional information, or clarifications to test takers during the test administration; test takers simply need to rely on their own test-taking skills.

Despite students' potential or likely familiarization with the format of the test and the type of tasks before the actual test is administered, it remains important that instructions for the tasks possess all the features discussed in this chapter for one chief reason: students' reading of instructions must be cultivated and honed through properly directed practice so as to prepare them for taking any form of a standardized test that they are sure to encounter at some point in their education and which will demand not merely knowledge but test-taking skills as well.

Chapter 4

Topics for discussion

1. Before reading this chapter, did you know that instructions are so impactful with respect to test-taking and test performance?

2. Look at the following task instructions and determine what test quality/qualities they impact (see Section 4.1).

(Grade 5) **Write the Past Simple of these verbs.**
Order –
Turn –
Drop –
Stay –
Carry –

(Grade 5) **The three forms of the verb TO BE in the Present Tense are:**

_____ _____ _____

(Grade 6) **Complete the sentences with possessive pronouns.**
I never wear clothes that aren't _____.
We bought the house last week. It's _____ now.
They can't sell the car. It's not _____.

(Grade 6) **Write the sentences in the Present Perfect.**
Steve / fly / in a plane. √
I / swim / with whales. x
George / climb / Mont Blanc. √
We / be / to Austria. x

(Grade 7) **Fill in the gaps!**
It _____ (not, work). I think it's broken.
_____ (it , rain) at the moment?
I _____ (have) lunch in the cafeteria every day.
Sheila _____ (love) reading books.
What _____ (you, be) up to?

(Grade 7) **Negate the sentences.**
They are using a dictionary.
She is good at sports.
Mum always puts an apple in my bag.

(Grade 8) **Complete the dialogue.**
Lyn Hi, Mark (1) is Debbie.
Mark Hi, Debbie. (2) to meet you!
Debbie Nice to (3) you, (4), Mark.

(Grade 8) **Fill in the blanks with relative pronouns WHICH, WHO, WHOSE:**
My brother, _____ is an engineer, helped me.
He chose the shoes _____ he wanted to buy.
The teacher _____ homework I never do rang my mom.
The boy _____ phoned didn't leave a message.

3. What features do the instructions for the tasks above have and/or lack (see Section 4.4)?

4. Rewrite the instructions for the tasks above so that they are more effective and contain all the necessary parts an instruction is suggested to have (see Section 4.3).

5. Could any of the instructions for the tasks above be contextualized/ personalized? If so, suggest how this could be achieved.

REFERENCES

Abbas, A. (1994). *Evaluating the assessment process in the EFL teaching programs and the General Secondary Education Certificate English Exams for 1989–1993*. [Master's thesis] Master's capstone projects, paper 118, University of Massachusetts Amherst. http://scholarworks.umass.edu/ cie_capstones/118.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing. A revision of Bloom's taxonomy of educational objectives* (complete edition). Longman.

Angelo, T, & Cross, K. P. (1993). *Classroom assessment techniques*. *A handbook for college teacher* (2nd edition). Jossey-Bass Publishers.

Alderson, C. J, Clapham, C., & Wall, D. (2002). *Language test construction and evaluation*. Cambridge University Press.

Assessment Reform Group. (2003). *The role of teachers in the assessment of learning.* CPA Office, Institute of Education.

Bachman, L. F. (1995). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (2004). *Language testing in practice: Designing and developing useful language tests.* Oxford University Press.

Beaumont, J. (2010). A sequence of critical thinking tasks. *TESOL Journal*, *1*(4), 427–448. https://doi. org/10.5054/tj.2010.234763

Bloom, B., Englehart, M., Furst, E., Hill, W., & Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Longmans.

Bobrowski, P. (2006). Bloom's taxonomy – expanding its meaning. *Faculty Guidebook*, 161–164. http://www.pcrest.com/research/fgb/2_2_1.pdf

Bodrič, R. (2016). Testiranje gramatike stranog jezika – savremena teorijska i praktična načela. *Nasleđe*, *34*, 159–170.

Boyle, J., & Fisher, S. (2007). *Educational testing. A competence-based approach*. Blackwell Publishing.

Brown, H. D. (2000). *Teaching by principles. An interactive approach to language pedagogy* (2nd edition). Pearson Education ESL.

Bruce, F., & Schmitt, V. (2010). Teachers' classroom assessment practices. *Middle Grades Research Journal*, *5*(3), 107–117.

Cohen, A. (1994). *Assessing language ability in the classroom* (2nd ed.). Heinle & Heinle Publishers.

Cohen, M., Salas, E., & Riedel, S. L. (2002). *Critical thinking: Challenges, possibilities, and purpose*. Cognitive Technologies, Inc.

Council of Europe (2002). *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press.

DiDonato, N., Fives, H., & Krause, E. (2013). Using a table of specifications to improve teacher-constructed traditional tests: An experimental design. *Assessment in Education: Principles, Policy & Practice, 21*(1), 90–108. https://doi.org/10.1080/0969594X.2013.808173

Dimitrijević, N. (1999). *Testiranje u nastavi stranih jezika* [*Testing in foreign language teaching*]. Zavod za udžbenike i nastavna sredstva.

Douglas, D. (2010). *Understanding language testing*. Routledge.

Facione, P. A. (1990). *Critical thinking: A statement of expert consensus for purposes of educational assessment and instruction*. The California Academic Press.

Fahim, M, Bagherkazemi, M., & Alemi, M. (2010). The relationship between test takers' critical thinking ability and their performance on the reading section of TOEFL. *Journal of Language Teaching and Research*, *1*(6), 830–837. https://doi.org/10.4304/jltr.1.6.830–837

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research and Evaluation*, *18*(3), 1–7. https://doi.org/10.7275/cztt-7109

Frey, B., Petersen, S., Edwards, L., Teramoto Pedrotti, J., & Peyton, V. (2005). Item-writing rules: Collective wisdom. *Teaching and Teacher Education*, *21*(4), 357–364. https://doi.org/10.1016/j.tate.2005.01.008

Frey, B., & Schmitt, V. (2007). Coming to terms with classroom assessment. *Journal of Advanced Academics*, *18*(3), 402–423. https://doi.org/10.4219/jaa-2007-495

Frey, B., & Schmitt, V. (2010). Teachers' classroom assessment practices. *Middle Grades Teachers' Research Journal*, *5*(3), 107–117.

Fulcher, G. (2010). *Practical language testing*. Hodder Education.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment. An advance resource book*. Routledge.

Glaser, R. (1984). Education and thinking: The role of knowledge. *American Psychologist*, *34*(2), 93–104.

Glušac, T., & Milić, M. (2021). Quality of written instructions in teacher-made tests of English as a foreign language. *English Teaching & Learning*, online, 1–19. DOI 10.1007/s42321-021-00079-1

Glušac, T., Milić, M., & Lužajić, D. (2021, November 20). *Test task instruction contextualization as a means of enhancing test performance and motivation* [Conference presentation]. Languages and Cultures in Times and Space 10, Novi Sad, Serbia. http://www.ff.uns.ac.rs/uploads/files/Nauka/Konferencije/2021/JiKuViP%2010/jikuvip10_prog-final.pdf

Glušac, T., & Pilipović, V. (2016). Developing critical thinking in teaching EFL through asking questions. In B. Vujin, & M. Radin-Sabadoš (Eds.), *English Studies Today: Words and Visions*. *Selected papers from The Third International Conference English Language and Anglophone Literatures Today (ELALT 3)* (pp. 401–415). Faculty of Philosophy, University of Novi Sad.

Glušac, T., & Pilipović, V. (2017). Značaj nastavničkih testova u nastavi stranih jezika [The importance of teacher-made tests in foreign language teaching]. *Nasleđe*, *36*, 285–297.

Glušac, T., Pilipović, V., & Marčićev, N. (2019). Analysis of English language test tasks for seventh- and eighth-graders in Serbia according to Bloom's Taxonomy. *Nastava i vaspitanje*, *68*(1), 35–50. DOI:10.5937/ nasvas1901035G

Glušac, T., Pilipović, V., & Milić, M. (2020). Critical thinking skills of third grade secondary school students as a component of their functional literacy. *Teme*, *29*(2), 355–380. https://doi.org/10.22190/TEME180206031G

Guskey, T. R. (2003). How classroom assessments improve learning. *Educational Leadership*, *6*(5), 6–11.

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking* (4th edition). Lawrence Erlbaum Associates, Inc. Publishers.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. Dispositions, skills, structure training, and metacognitive monitoring. *American Psychologist*, *53*(4), 449–455. https://doi.org/10.1037/0003-066X.53.4.449

Hamp-Lyons, L. (2016). The purpose of assessment. In D. Tsagardi, & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 13–27). DeGruyter Mouton.

Hatipoğlu, C. (2015). English language testing and evaluation (ELTE) training in Turkey: Expectations and needs of pre-service English language teachers. *ELT Research Journal*, *4*(2), 111–128.

Hattie, J., & Brown, G. (2010). Assessment and evaluation. In C. Rubie-Davies (Ed.), *Educational psychology: Concepts, research and challenges* (pp. 102–117). Routledge.

Heaton, J. B. (1990). *Writing English language tests*. Longman Group UK Limited.

Hidri, S. (Ed.) (2021). *Perspectives on language assessment literacy. Challenges for improved student learning*. Routledge, Taylor & Francis Group.

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

Lai, E. R. (2011). *Critical thinking: A literature review. Research Report*. Pearson.

Lakin, J. (2014). Test directions as a critical component of test design: Best practices and the impact of examinee characteristics. *Educational Assessment*, *19*(1), 17–34. https://doi.org/10.1080/10627197.2014.869448.

Marso, R. N., & Pigge, F. L. (1988, April 6-8). *An analysis of teacher-made tests: testing practices, cognitive demands, and item construction errors* [Conference presentation]. Annual meeting of the National Council on Measurement in Education, New Orleans, Louisiana.

Marso, R. N., & Pigge, F. L. (1991). An analysis of teacher-made tests: Testing practices, cognitive demands and item construction errors. *Contemporary Educational Psychology*, *16*, 279–286.

Marso, R. N., & Piggie, F. L. (1993). Teachers' testing knowledge, skills and practices. In S. Wise (Ed.), *Teacher training in measurement and assessment skills* (pp. 129–185). Buros Institute of Mental Measurements, University of Nebraska–Lincoln. https://digitalcommons.unl.edu/cgi/viewcontent.cgi?article=1007&context=burosteachertraining

Martínez, J. F., B. Stecher, B., & Borko, H. (2009). Classroom assessment practices, teacher judgment, and student achievement in mathematics:

Evidence from ECLS. *Educational Assessment*, *14*(2), 78–102. https://doi.org/10.1080/10627190903039429

McKey, P. (2008). *Assessing young language learners*. Cambridge University Press.

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research & Education*, *7*(8), 1–9. https://doi.org/10.7275/5kc4-jy05

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research, & Evaluation*, *7*, 1–5. DOI: https://doi.org/10.7275/5kc4-jy05

McMillan, J. H., Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research*, *95*(4), 203–213. https://doi.org/10.1080/00220670209596593

McPeck, J. E. (1981). *Critical thinking and education*. St. Martins Press.

Mirkov, S., & Gutvajn, N. (2014). How students perceive teachers' activities aimed at stimulating critical thinking. *Nastava i vaspitanje*, *63*(4), 621–636.

Mirkov, S., & Stokanić, D. (2015). Podsticanje kritičkog mišljenja kod učenika: stavovi i aktivnosti nastavnika [Promoting students' critical thinking: Teachers' attitudes and activities]. *Inovacije u nastavi*, *28*(1), 25–41. https://doi.org/10.5937/inovacije1501025M

Morais F., Silva, H., Cruz, G., Pedrosa, D., Payan-Carreira, R., Dominguez, C., & Nascimento, M. M. (2019). Perceptions of Portuguese university teachers about critical thinking educational practices. In M. Tsitouridou, A. Diniz, & T. Mikropoulos (Eds), *Technology and Innovation in Learning, Teaching and Education, TECH-EDU 2018, Communications in Computer and Information Science*, 993 (pp. 223–239). DOI.org/10.1007/978-3-030-20954-4_17.

Paul, R., & Elder, L. (2008). Critical thinking: The nuts and bolts of education. *Optometric Education*, *33*(3), 88–91. https://journal.opted.org/files/Volume_33_Number_3_Summer_2008.pdf

Pešić, J. (2011). Sličnosti i razlike u konceptualizovanju kritičkog mišljenja [Similarities and differences in conceptualizing critical thinking]. *Psihološka istraživanja*, *14*(1), 5–23.

Pravilnik o nastavnom planu za drugi ciklus osnovnog obrazovanja i vaspitanja i nastavnom programu za peti razred osnovnog obrazovanja i vaspitanja [Rulebook on the Syllabus for the Second Cycle of Primary School Education and Curriculum for the Fifth Grade of Primary School]. *Službeni glasnik RS – Prosvetni glasnik*, br. 6/2007, 2/2010, 7/2010 − dr. pravilnik, 1/2013, 5/2014 i 11/2016.

Pravilnik o nastavnom planu za drugi ciklus osnovnog obrazovanja i vaspitanja i nastavnom programu za šesti razred osnovnog obrazovanja i vaspitanja [Rulebook on the Syllabus for the Second Cycle of Primary School Education and Curriculum for the Sixth Grade of Primary School]. *Službeni glasnik RS – Prosvetni glasnik*, br. 5/2008, 3/2011, – dr. pravilnik 3/2011 – dr. pravilnik, 1/2013, 4/2013, 11/2016 i 6/2017.

Pravilnik o nastavnom programu za sedmi razred osnovnog obrazovanja i vaspitanja [Rulebook on the Curriculum for the Seventh Grade of Primary School]. *Službeni glasnik RS – Prosvetni glasnik*, br. 6/2009, 3/2011 – dr. Pravilnik, 8/2013, 11/2016 i 12/2018.

Pravilnik o nastavnom programu za osmi razred osnovnog obrazovanja i vaspitanja [Rulebook on the Curriculum for the Eighth Grade of Primary School]. *Službeni glasnik RS – Prosvetni glasnik*, br. 2/2010, 3/2011 – dr. Pravilnik, 8/2013, 5/2014, 11/2016 i 12/2018.

Prošić-Santovac, D., Savić, V., & Rixon, S. (2019). Assessing young English language learners in Serbia: Teachers' attitudes and practices. In D. Prošić-Santovac & S. Rixon (Eds.), *Integrating assessment into early language learning and teaching* (pp. 251–265). Multilingual Matters. https://doi.org/10.21832/ 9781788924825-019

Purpura, J. (2004). *Assessing grammar*. Cambridge University Press.

Radić-Bojanić, B., & Topalov, J. (2016). Textbooks in the EFL classroom: Defining, assessing, analyzing. *Collection of Papers of the Faculty of Philosophy of Priština*, *XLVI*(3), 137–153.

Rubin, R. (2017, May 18). *Developing your students' vocabulary and grammar for critical thinking*. American English Webinar Series 6. Regional English Language Office of the US Embassy, Hanoi. [Video] https://www.youtube.com/watch?v=uUMB_ASX5Gs

Rudner, L., & W. Schafer, W. (2002). *What teachers need to know about assessment*. National Education Association.

Shepard, L. A. (2000). *The role of classroom assessment in teaching and learning.* Center for the study of evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies University of California, Center for Research on Education, Diversity and Excellence.

Starr, S. (2014). Moving from evaluation to assessment. *Journal of the Medical Library Association*, *102*(4), 227–229. doi: 10.3163/1536-5050.102.4.001

Tsagari, D., Vogt, K., Froehlich, V., Csepes, I., Fekete, A., Green, A., Hamp-Lyons, L., Sifakis, N., & Kordia, S. (2018). *Handbook of assessment for language teachers*. European Commission.

Wattles, I. (2016). *Visoke kognitivne funkcije u nastavi lingvističkih predmeta na tercijarnom nivou obrazovanja* [*Higher Cognitive Functions in Linguistic Courses in Tertiary Education*] [Doctoral dissertation]. Faculty of Philosophy, University of Novi Sad.

Weir, C. J. (1993). *Understanding and developing language tests*. Prentice Hall.

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

Weigle, S. C. (2009). *Assessing writing*. Cambridge University Press.

Wilson, K. (2016). Critical reading, critical thinking: Delicate scaffolding in English for academic purposes (EAP). *Cognitive Processes and Creativity*, *22*, 256–265. https://doi.org/10.1016/j.tsc.2016.10.002

Yanning, D. (2017). Teaching and assessing critical thinking in se cond language writing: An infusion approach. *Chinese Journal of Applied Linguistics*, *40*(4), 431–451. DOI:10.1515/cjal-2017-0025

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, *16*(4), 323–342. https://doi.org/10.1207/S15324818AME1604_4

Dr. Tatjana Glušac

NEW INSIGHTS INTO FOREIGN LANGUAGE TESTING

Publisher:
Faculty of Law and Business Studies Dr. Lazar Vrkatić*, Novi Sad*

Representing the Publisher:
Prof. Mirjana Franceško

Editor-in-chief:
Prof. Vladimir Njegomir

Reviewers:
Prof. Vesna Pilipović
Faculty of Law and Business Studies Dr. Lazar Vrkatić*, Novi Sad*

Prof. Radmila Bodrič
Faculty of Philosophy, Novi Sad

Dr. Mira Milić
Faculty of Sport and Physical Education, Novi Sad

Prof. Çiler Hatipoğlu
Faculty of Education, Middle East Technical University, Ankara, Turkey

Language adviser and proofreader:
Andrew Wiesike

Cover design:
Sonja Vrkatić

Preliminary typesetting:
Aleksandar Međedović

Typesetting and conversion:
Ferenc Finčur
KriMel, Budisava

ISBN 978-86-7910-154-9

2022

Prof. dr Tatjana Glušac

NEW INSIGHTS INTO FOREIGN LANGUAGE TESTING

Izdavač:
Fakultet za pravne i poslovne studije dr Lazar Vrkatić*, Novi Sad*

Za izdavača:
Prof. dr Mirjana Franceško

Glavni i odgovorni urednik:
Prof. dr Vladimir Njegomir

Recenzenti:
Prof. dr Vesna Pilipović
Fakultet za pravne i poslovne studije dr Lazar Vrkatić*, Novi Sad*

Prof. dr Radmila Bodrič
Filozofski fakultet, Novi Sad

Prof. dr Mira Milić
Fakultet sporta i fizičkog vaspitanja, Novi Sad

Prof. dr Çiler Hatipoğlu
Faculty of Education, Middle East Technical University, Ankara, Turska

Lektor i korektor:
Andrew Wiesike

Dizajn korica:
Sonja Vrkatić

Kompjuterska obrada rukopisa:
Aleksandar Međedović

Kompjuterska priprema i konverzija:
Ferenc Finčur
KriMel, Budisava

ISBN 978-86-7910-154-9

2022.

CIP - Каталогизација у публикацији
Библиотеке Матице српске, Нови Сад

37.091.26

GLUŠAC, Tatjana, 1976-
New insights into foreign language testing [Elektronski izvor] / Tatjana Glušac. - Novi Sad : Faculty of Law and Business Studies Dr. Lazar Vrkatić, 2022

Način pristupa (URL): http://www.flv.edu.rs/izdavastvo. - Opis zasnovan na stanju na dan 17.1.2022. - Nasl. s naslovnog ekrana. - Bibliografija.

ISBN 978-86-7910-154-9

a) Тестови знања -- Страни језици

COBISS.SR-ID 55996681

Beautifully, clearly and systematically written, this book focuses on a number of carefully selected topics in the area of foreign language testing that have not received enough attention thus far. After discussing the basic concepts of foreign language testing, the author engages in a deeper analysis of commercial and teacher-made tests, levels of cognitive processing in test item design, and the qualities of, as well as frequent oversights in, written test instructions, supporting her convincing arguments with numerous examples from teaching practice, relevant research results, and prominent and up-to-date literature resources. This is why this book is both of scientific and practical value - being equally useful as a resource in ELT academic research and as a manual in teacher training courses.

Prof. Vesna Pilipović,
Faculty of Law and Business Studies dr Lazar Vrkatić*, Novi Sad*

The publication NEW INSIGHTS INTO FOREIGN LANGUAGE TESTING is the result of the author's extensive and versatile L2 teaching and testing experience and research, continuous professional development and critical reflections, driving her to competently explore, discuss and offer informed L2 testing practices and valid arguments for the preferred use of teacher-made tests over ready-made ones. This well-conceived and successfully written book is committed to the most important theoretical and practical aspects of foreign language testing and assessment for learning. As such it represents a very useful and rich repository of knowledge, ideas and advice for L2 MA and BA students, novice and experienced practising teachers and scholars alike.

Prof. Radmila Bodrič,
Faculty of Philosophy, Novi Sad

The Book explores the practices in foreign language assessment, focusing on teacher-made tests. From the user's perspective, it is a multi-purpose resource. First of all, it contributes to research findings of scholarly interest. Next, it provides a teaching resource at the tertiary level. Finally, it might serve as a practical guide for foreign language teachers that have not been trained in this field. Thus, the topic of high significance, sound argumentation based on long-year research and experience of the author, as well as well-written and neatly-structured reading material, offers valuable content to all those dealing with theoretical, practical, and educational issues of foreign language assessment.

Dr. Mira Milić,
Faculty of Sport and Physical Education, Novi Sad

New Insights into Foreign Language Testing is a fresh and exciting addition to the existing work in the foreign language field. The author should be commended for her meticulous and innovative examination of a topic (i.e., foreign language testing) that has a real bearing on the lives of all who learn and take foreign language exams. It is an impressive and stimulating read that successfully balances theory, ideas, and practical exercises in each chapter.

Prof. Çiler Hatipoğlu,
Faculty of Education, Middle East Technical University, Turkey