Diagnostic Tests in Czech for the Children of Immigrants Attending Primary Schools

Kateřina Vodičková, Yvona Kostelecká

Abstract

Mastering the second language, Czech in our case, is crucial for children of migrants, so that among other things they can integrate into the education process. To adjust the language teaching to the pupils' needs, it is necessary to identify what language skills, or individual competences within the frame of the communicative competence, should be developed. For this purpose, a new diagnostic test for the lower graders and upper graders of primary schools was designed. Although it is not a high-stakes test, it is essential to ensure its validity, reliability, practicality as well as its positive impact on the teaching process, the pupils, their teachers and the society. This paper presents the diagnostic tool, documents the qualities of the measurement device and outlines its further development.

Key words: Common European Framework of Reference for Languages (CEFR), Czech for foreigners, diagnostic test, language testing, young learners

Abstrakt

Pro žáky-cizince je klíčové ovládnutí druhého jazyka, v našem případě češtiny, mimo jiné proto, aby se mohli začlenit do procesu vzdělávání. Aby bylo možné zacílit jazykovou výuku podle potřeb žáků, je nutné určit, které řečové dovednosti, resp. které dílčí

kompetence v rámci komunikační kompetence je záhodno rozvíjet. Za tímto účelem byl vyvinut diagnostický test pro 1. a 2. stupeň základních škol. Ačkoliv se nejedná o test tzv. vysoké důležitosti, je zásadní, aby se jednalo o nástroj validní, reliabilní a praktický, aby měl pozitivní vliv na výuku, na žáky-cizince, na jejich učitele i na společnost. V tomto článku diagnostický nástroj představujeme, dokládáme jeho kvality a nastiňujeme další vývoj testu.

Klíčová slova: čeština jako druhý jazyk, diagnostický test, jazykové testování, Společný evropský referenční rámec pro jazyky (SERRJ), žák-cizinec.

Introduction

The number of children who attend primary schools and whose language of schooling differs from their first language (L1) has been growing steadily in the Czech Republic as well as in most European countries.[1] This growth has increased the interest not only in second-language (L2) instruction for young learners[2]/children of immigrants, but, in the first place, in questions like what level of communicative competence they have in Czech, what their progress is in L2, and how their (not only language) integration could be assisted. Students with insufficient Czech language skills may suffer feelings of social isolation and be ostracised by others.

With the support of Czech Science Foundation (project entitled Integration of the children of non-nationals into the Czech elementary schools, registration number: 13-32373S), the diagnostic test for the lower graders and upper graders in Czech primary schools was designed to map the children of immigrants' knowledge of Czech first generally, second individually. The findings as for the general tendencies are presented in e. g. Kostelecká et al. (2014), Kostelecká and Jančařík (2014) and are not the subject of this paper. As for the individual outcomes, there was a general hope that thanks to these tests, it would be possible to measure the pupils' progress in L2, and that the results would help the pupils, the teachers and the school to identify the fields in which the pupils need further instruction.

This paper introduces a set of diagnostic tests for children of immigrants, documents the qualities of the tests and suggests further steps to be taken for improving the tests. Before describing the format of the tests in Part 3, diagnostic tests are defined and

[1] In 1985, there were as few as 37,000 foreign nationals dwelling on the territory of what is now the Czech Republic (approximately 0.36 % of the total population), whereas in 2009, foreign nationals made up more than 4 % of the total population. In the 2010/2011 academic year, 1.7 % children enrolled in primary and lower secondary schools were children of immigrants.

[2] The term "young learners" in literature usually refers to children aged between 6 or 7 to about 12. Since it does not cover the ages of all children attending primary schools in the Czech Republic, the term "children of immigrants" is preferred in this paper.

characterised generally in Part 2. Part 4 deals with assuring the quality of the tests and Part 5 summarises steps to be taken to ensure further improvements of the diagnostic tests. Concluding remarks can be found in Part 6.

1 Diagnostic testing

Generally, diagnostic tests should guide teachers and/or schools and other institutions as well as pupils or students; the results of these tests are meant to help teachers and schools to adjust the instruction, provide learners with relevant feedback and indicate where they need to improve themselves and, ideally, suggest how this aim could be reached. A diagnostic test is thus seen as a test "which is used for the purpose of discovering a learner's specific strengths or weaknesses. The results may be used in making decisions on future training, learning or teaching." (ALTE 1998, p. 142) Similarly, Alderson, Clapham and Wall (1995, p. 12) state the purpose of diagnostic tests as "to identify those areas in which a student needs further help".

Despite the general consent on the purpose of diagnostic tests and the use of the results, their definitions vary. In literature, diagnostic tests are usually distinguished from other types of tests such as placement, progress, achievement, and proficiency tests. However, references to the common nature of diagnostic tests and especially placement or proficiency tests are not rare (cf. e. g. Davies et al., 1999, p. 43). Bachman (1990, p. 60) argues that "virtually any language test has some potential for providing diagnostic information". Alderson (1995, p. 12) clearly states that "achievement and proficiency tests are themselves frequently used … for diagnostic purposes". Bachman (1990, p. 60) points out that diagnostic tests may be either syllabus-based, which supports the aforementioned close relation to achievement tests, or theory-based.

Theory-based diagnostic tests tend to be proficiency tests if based on models of communicative language ability or communicative competence. Hughes (2003) also suggests that proficiency tests may serve as diagnostic tests, but it depends on the exact purpose of diagnosis. Likewise, Alderson, Clapham and Wall (1995, p. 12) claim that diagnostic tests "can be fairly general, and show, for example, whether a student needs particular help with one of the four main language skills". Furthermore, these "general", i. e. skill-based, tests offer the possibility of more detailed analysis of written and spoken performance and consequently enable the researchers to focus on the individual components of the linguistic competence (e. g. orthographical and grammatical competence) or on the specific phenomena (e. g. in pronunciation).

So what makes a test diagnostic rather than placement, achievement or proficiency? The answer seems to lie in the interpretation and use of the results in the first place. Alderson (2005, p. 10) emphasizes that there is a lack of guidance on designing the

diagnostic tests, their possible content, and underlying theoretical basis in literature. Nonetheless, he summarises the characteristics a diagnostic test should have or usually demonstrates. These include, among others, the ability to identify learners' strengths and weaknesses, leading to remediation in further instruction, making a detailed analysis and report possible, providing immediate results, being low-stakes.

2 Diagnostic tests for young learners in Czech

One of the first diagnostic tests focusing on language skills in Czech were developed and described in 2007 (Cvejnová et al. 2007). However, these tests were designed for adult learners. A suite of diagnostic tests for children of immigrants was developed in the course of 2010–2014. Using an existent placement, achievement or proficiency test was not considered appropriate, mainly for the following reasons: a) There is a lack of Czech language tests designed exclusively for young learners and, to our best knowledge, none for children of immigrants. b) Using syllabus-based achievement tests, even if they existed, would not take into account the fact that the children may have learned Czech from various sources or from no official teaching materials at all. There is no specific syllabus that has to be covered before the test and that should be covered afterwards. c) The proficiency test Czech Language Certificate Examination for Young Learners (CCE-A1 for Young Learners and CCE-A2 for Young Learners[3]) is first, subject to a fee; second, available only at A1 and A2 levels according to the Common European Framework of Reference for Languages (2001, hereafter CEFR); third, too time-consuming.

For these reasons, a pilot version of a tailor-made diagnostic test for primary schools was introduced in 2010. The test was decided to be a proficiency test as there is no syllabus the test can be related to. For this reason, there is also no grammar or vocabulary test although some information on the level of grammatical, lexical and other competencies can be inferred from the productive-skills subtests.

2.1 General overview of the diagnostic tests

Within the project no. 13-32373S of the Czech Science Foundation, two diagnostic tests were developed. One of them is aimed at lower graders. Taking in consideration the development of language skills in L1 and the cognitive development of the respondents, this test is designed for pupils attending the 3rd, 4th and 5th grades, which roughly corresponds to the ages from 8 to 11. It verifies the level of communicative competence within the language skills at A1 and A2 levels according to the CEFR. The other test is

[3] http://ujop.cuni.cz/cce-mladez

aimed at upper graders, i. e. the age group between 12 and 16, and verifies the level of language skills at A1, A2 and B1 levels according to the CEFR.

When designing the test, the authors could not base the test directly on the CEFR and its descriptors, since these are defined for adult language users and they do not take into account the children's cognitive development and communicative situations they enter. Therefore the tests are founded on the documents based on the CEFR, namely the language portfolios – the diagnostic test for lower graders is based on the Portfolio for Learners Up to the Age of 11 Vacková et al., 2002), the one for upper graders is based on European Language Portfolio for Learners aged 11 to 15 (Perclová, 2010).

2.2 The format of the diagnostic test for lower graders

The lower-grader diagnostic test at A1 and A2 levels verifies all four language skills in four subtests: reading, listening, writing, and speaking. The pupils can gain the maximum of 15 points in each subtest per level (see Table 1).

Table 1
The format of the lower-grader diagnostic test

Level	Subtest	No. of tasks	No. of points	Time
A1	Listening	3	5+5+5	10 minutes
	Reading	3	5+5+5	12 minutes
	Writing	2	6+9	10 minutes
	Speaking	1	15	3 minutes
A2	Listening	3	5+5+5	15 minutes
	Reading	3	5+5+5	18 minutes
	Writing	2	6+9	15 minutes
	Speaking	1	15	5 minutes

2.3 The format of the diagnostic test for upper graders

The upper-grader diagnostic test verifies the level of communicative competence in four language skills at A1, A2 and B1 levels according to the CEFR. The format of the test corresponds to the format of the diagnostic test for lower graders (cf. Table 2) although the test techniques may vary and so does the time allotted to each subtest. It should be noted that there is only one task in the subtest Writing at A2 and at B1 level to eliminate the error rate caused by fatigue and reduced concentration.

Table 2
The format of the upper-grader diagnostic test

Level	Subtest	No. of tasks	No. of points	Time
A1	Listening	3	5+5+5	6 minutes
	Reading	3	5+5+5	10 minutes
	Writing	2	5+10	10 minutes
	Speaking	1	15	3 minutes
A2	Listening	3	5+5+5	9 minutes
	Reading	3	5+5+5	10 minutes
	Writing	1	15	10 minutes
	Speaking	1	15	4–5 minutes
B1	Listening	2	5+10	13 minutes
	Reading	3	5+5+5	15 minutes
	Writing	1	15	15 minutes
	Speaking	1	15	4–5 minutes

3 Qualities of the diagnostic tests

A number of guidelines is available that help the test designers build the validity argument. These include e. g. ALTE Minimum standards for establishing quality profiles[4] and EALTA Guidelines for good practice in language testing and assessment[5]. In this paper, we follow the ALTE Minimum standards to support the validity argument.

3.1 Test construction

As for the theoretical model, the test is based on the Bachman and Palmer (1996) framework of language ability, which is widely used in language assessment and at the same time it is a model used in assessing young learners by e. g. McKay (2008, p. 51). The model ensures that the language can be tested in communication, within four language skills, and that general language can be tested as well. The components of the model have been taken into account when designing the test, test specifications, and assessment criteria. The model is operationalized via test specifications that state what the pupils need to know in order to fulfil the given communication goals in a particular way and, thus, to reach a particular level of communicative competence in the examination.

[4] http://www.alte.org/attachments/files/minimum_standards.pdf

[5] http://www.ealta.eu.org/documents/archive/guidelines/English.pdf

The test specifications are based on the aforementioned European Language Portfolios, thus ensuring the communicative goals, sub-skills, text types, domains, etc. correspond to the CEFR, or more precisely, to the documents based on the CEFR.[6] Although there is no need for a large number of test versions because of the purpose of the test (pupils take the test once in their live or they re-take the test after a longer time), the test versions are comparable regarding their content thanks to the detailed test specifications.

A team of test constructors was recruited among experienced test constructors and item writers working for the Institute for Language and Preparatory Studies, Charles University in Prague, who are familiar with teaching and/or testing young learners. For the reviews and revisions, an internal expert from Charles University was selected as well as some external reviewers who are experienced in language testing and/or teaching young learners.

3.2 Administration and logistics

No administration centres are needed, as the pupils sit the test in the primary school they attend. Since the numbers of pupils taking the test have not been vast, so far the test has been administered, the Speaking part examined, and all subtests rated by the test constructors. Thus, all training and benchmarking take rather the form of team consultations and meetings.

3.3 Marking and grading

All subtests are marked by the team of administrators/examiners. The Reading and Listening parts are corrected according to a key, which does not require any special training. Since the team is quite small and consists of test constructors, the training for raters of productive skills took a form of discussions and rating sample performances. Regarding the receptive and productive skills, random checks of the accuracy of marking are introduced.

3.4 Test analysis

The piloting phase took place throughout the year 2010 using the first version of the test. After revisions based on the results and experience from the piloting, pretesting took place under the same test conditions in 2013. To ensure that both the piloted and pretested population is the same as the intended test population, the piloting and

[6] The external reviewers appreciated that the link to the external framework of reference was evident and the difficulty of the test corresponds to the levels of communicative competence.

pretesting were realised on a voluntary basis at a number of primary schools. Only the children with different L2 from Czech from 3rd to 9th grades were allowed to take the test if their parents approved.

When interpreting the results from the piloting stage, mainly qualitative analysis was used and interviews with some pretested participants were carried out. Quantitative analysis (item analysis) was used to calculate especially facility values and the discrimination index of the items. The analysed data was used to verify how well the tasks function. After revisions, pretesting was arranged.

For the test analysis, two different theories are used. One is the Classical Test Theory (Lord, Novick 1968; Crocker, Algina 1986), the other is the Item Response Theory, published by Rasch in 1960. For the data from pretesting, we used statistical software Iteman 4.1 based on the Classical Test Theory for both diagnostic tests. In the case of lower-grader diagnostic test comprising A1 and A2 levels, Reading and Listening at both levels and the first A1 task in Writing were analysed. The test was taken by 129 respondents. In the case of the upper-grader diagnostic test comprising A1, A2 and B1 levels, Reading and Listening at all levels and the first A1 task in Writing were analysed. The test was taken by 132 respondents.

3.4.1 Reliability of the diagnostic tests

Řehák (1998) understands reliability as the accuracy of measuring the characteristics we measure in reality and Kreidl (2004) defines it as the accuracy, consistency of measurement, i. e. the ability to reach the same result of measurement in case the state of the observed object has not change. Similarly, Chráska (2007) states that a test holds high reliability if the results are trustworthy and accurate. In his opinion, results are trustworthy if the same or very similar values are acquired under constant test conditions and accurate if the influence of errors on the results is minimised.

Test reliability cannot be measured accurately; it is only estimated and reported through the reliability coefficient in practice. The closer the coefficient comes to +1, the more accurate and reliable the test is (Schindler, 2006).

Nonetheless, Soukup (2005) adds that reliability of a test should be interpreted with caution as it depends on the number of the items. The more items, even if useless, appear in the test, the higher value of the reliability coefficient. Chráska (2007) claims that generally, a reliability coefficient of 0.8 and above is considered optimal and 0.95 even excellent for didactic tests.

Reliability of a test can be estimated in two ways – by parallel measurements (Test-Retest method, Parallel-Forms method) or by internal consistency (splitting the test in two halves and estimating the internal consistency). For the Test-Retest method, it is necessary to re-take the test after a certain period of time. This method was considered unfeasible in the case of the aforementioned diagnostic tests. Using parallel tests was

not considered practical either, as there would have to be two parallel versions of the test and pupils would have to take both of them.

The most frequently used method of estimating reliability is the internal consistency method which can be applied only in test with homogenous content. This method presupposes that the answers to all items measuring the same characteristics hold sufficiently high positive correlation. For calculating the inner consistency of a test, Cronbach's alpha (Cronbach 1951) is probably the most widely used formula:

$$\alpha = kk - 11 - i = 1k\sigma i2\sigma t2$$

where *k* is the number of items, σi2 is the variance of component *i* for the current sample of respondents and the σt2 is the variance of the observed total test scores. Cronbach's alpha comes from Kuder-Richardson method (Kuder, Richardson 1937), or more precisely from Kuder-Richardson Formula 20 (KR-20); it made it possible to estimate reliability for multiple choice items as well. Kuder-Richardson method can be seen as a specific case of Cronbach's alpha; it is used for dichotomous items and the reliability coefficient is calculated from the following formula:

$$\alpha = kk - 11 - i = 1kpiqi\sigma t2$$

where *k* is the number of items, pi is the proportion of correct responses to test item *i*, qi = 1 – pi is the proportion of incorrect responses to test item *i* and σt2 is the variance of the observed total test scores.

Another method used for estimating reliability of a test is the Split-Half method. Chráska (2007) considers this method appropriate for a test with items ordered according to their difficulty from the easiest ones to the most difficult ones. This method presupposes that if the test is reliable, its parts, namely two halves, must be reliable, too. These halves are assessed separately and then the results are correlated. The correlation between the two halves is corrected by Spearman-Braun Formula (Chráska 2007):

$$rab = 2rp1 + rp$$

where rab is the reliability coefficient and rp is the reliability coefficient between the results of both halves of the test.

Table 3 shows the reliability coefficients gained by applying Kuder-Richardson Formula in the lower-grader test as a whole as well as in its two parts. On top of that, it also shows the reliability coefficient gained by the Split Half method in three variants of halving the set: Split-Half Random (items are split into halves at random), Split Half First-Last (one set consists of the first half of the items, the other set of the second half), and Split Half Odd Even (one set comprises the odd items, the other one the even items).

For all variants of splitting, results are shown for both non-corrected variant and the variant corrected by Spearman-Braun Formula. This correction is used because in the non-corrected version we compare two tests with only half of the items that the live test has. Standard error of measurement (SEM), which estimates the standard deviation of the errors of measurement in the scale scores, is reported, too.

Table 3
Reliability coefficients for the lower-grader diagnostic test

	Alfa (KR-20)	SEM	Non-corrected			Spearman-Braun Correction		
			Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)
Whole test	0.951	2.180	0.921	0.796	0.919	0.959	0.886	0.958
A1 test	0.915	1.373	0.844	0.755	0.878	0.915	0.860	0.935
A2 test	0.915	1.663	0.804	0.794	0.867	0.891	0.885	0.929

The data in Table 3 show that when applying the Kuder-Richards formula, the reliability coefficient exceeds 0.9 for the individual tests and reaches even 0.95 for the whole test. Slightly lower reliability coefficients occur when using the split-half method. However, it should be noted that splitting a diagnostic test in two equivalent halves is complicated. Since the tasks and items are ordered according to their difficulty, we get the lowest reliability coefficient when comparing the first and the second half of items (Split-Half First-Last Method). The reliability coefficient is considerably higher when the Random or Odd-Even variant of the Split-Half method is used. In these cases, it almost always exceeds 0.9.

Similarly to Table 3, Table 4 shows the same test characteristics for the upper-grader diagnostic test.

Table 4
Reliability coefficients for the upper-grader diagnostic test

	Alfa (KR-20)	SEM	Non-corrected			Spearman-Braun Correction		
			Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)
Whole test	0.971	2.523	0.904	0.798	0.952	0.949	0.888	0.976
A1 test	0.920	1.319	0.825	0.697	0.878	0.904	0.821	0.935
A2 test	0.944	1.190	0.898	0.769	0.921	0.946	0.869	0.959
B1 test	0.934	1.663	0.881	0.805	0.885	0.937	0.892	0.939

In this case, when Kuder-Richardson Formula is applied the reliability coefficients are even higher than in the case of the lower-grader test. High values of reliability coefficient are gained also when using the Split-Half method.

3.4.2 Difficulty and discrimination

Another characteristics used for judging the tasks and items is the item difficulty and discrimination. It is generally agreed that the items of a didactic test should be neither too difficult nor too easy. The value of the difficulty index is between 0 and 1. Items with $P < 0.3$ are considered very difficult, whereas items with $P > 0.89$ are easy and, consequently, have low discrimination index.

According to Chráska (2007), discrimination refers to the ability of the test to distinguish between test takers that are more competent and less competent. The discrimination index is calculated in a number of ways and can acquire values from –1 to +1. The higher the discrimination index, the better the test distinguishes test takers with higher knowledge from test takers with lesser knowledge. In accordance to Ebel (1954), the discrimination index over 0.2 is generally considered sufficient. For the analysis of the diagnostic tests, the point-biserial coefficient (Rpbis) was used. The advantage is that Rpbis takes into account the difficulty of the item (Bílek, Jeřábek, 2010).

The values of the difficulty index and discrimination index can be found in Tables 5 and 6.

Table 5
Characteristics of the lower-grader diagnostic test

	No. of items	Average raw score	Standard deviation	Minimum score	Maximum score	Mean P Difficulty index	Mean Rpbis Discrimination
Whole test	72	64.419	9.887	7	72	0.895	0.498
A1 test	36	33.093	4.721	7	36	0.919	0.480
A2 test	36	31.326	5.707	0	36	0.870	0.517

The item analysis of the diagnostic test for lower graders shows that the average difficulty of items in both parts is relatively low (the average difficulty index is higher than 0.8 in both parts). Nonetheless, A1 test is easier than A2 test, which is positive, because within a diagnostic test, we expect that the higher the level of the test, the more difficult the test. At the same time, relatively low difficulty does not matter to a larger extent within a diagnostic test if the discrimination values are reasonably high. The whole test as well as the tests at both levels have a rather high discrimination – Rpbis is exceeding 0.45. For this reason, we believe the lower difficulty of the test is acceptable.

Table 6
Characteristics of the upper-grader diagnostic test

	No. of items	Average raw score	Standard deviation	Minimum score	Maximum score	Mean P Difficulty index	Mean Rpbis Discrimination
Whole test	95	80.598	14.777	0	91	0.848	0.532
A1 test	35	29.348	4.664	0	32	0.839	0.460
A2 test	30	26.424	5.016	0	29	0.881	0.595
B1 test	30	24.826	6.484	0	30	0.828	0.552

The results for the diagnostic test for upper graders are similar although the average value of difficulty index is slightly lower than in the test for lower graders. The discrimination index within all three parts of the test (A1, A2, B1) is again high.

4 Further steps

Part 4 shows not only what has been done in diagnostic testing within this project, but it also identifies fields in which further development is desirable.

First, it will be necessary to train the administrators, examiners and possibly also raters if the numbers of test takers grow. In the piloting and pretesting phase, these roles could have been handled by the team of test constructors since the number of test takers was relatively low. If the test is used on the national level (although probably voluntarily), more staff will be required to participate in test administration, examination and assessment. In diagnostic testing, prompt and detailed feedback both to test takers and teachers or schools is crucial. With growing numbers of test takers, it may be also necessary to train a number of experts in providing feedback to the test users.

Second, the team may need to focus on adapting the test to test takers with special needs.

Third, organising a standard setting in order to set the cut-off score may be useful. Although to our best knowledge, the number of even high-stakes tests in Czech that have gone through standard setting when establishing the cut-off score is extremely limited and in most cases, if not all, the cut-off score has not been implemented yet, running such a procedure would increase the quality of the test. On the other hand, it should be noted that the nature of the test makes a detailed report of e. g. subskills possible, which is probably more desirable than reporting a mere number or pass/fail result.

Fourth, some spoken and written performances could be double marked. This would allow for tracing the inter-rater reliability. Double marking would be important especially when the number of raters increases since it can help the test constructors

identify inconsistent raters, but also those that are too harsh or too lenient. Based on the findings, some raters may need to be re-trained and/or supervised.

Concluding remarks

The educational integration of students who are not native speakers of Czech is a subject that, given the rapid increase in immigration to the Czech Republic in the past twenty years, is a very relevant issue today. Since the number of immigrants in the Czech population is likely to grow even more, its relevance will only increase in the future.

The diagnostic test for lower graders and upper graders at Czech primary schools whose L1 is different from the language of instruction represents one of the first attempts to design an instrument that would help teachers, schools, and children of migrants with (language) integration. The paper presents the qualities of the test as well as areas that require further development.

References

Alderson, J. Ch. (2005). *Diagnosing Foreign Language Proficiency: the Interface Between Learning and Assessment*. London: Continuum.

Alderson, J. Charles, Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

University of Cambridge., & Association of Language Testers in Europe. (1999). *Multilingual glossary of language testing terms*. Cambridge: Cambridge University Press.

Bachman, L. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Council of Europe. (2001). *Common European framework of reference for languages: Reference, teaching, assessment*. Cambridge: Cambridge University Press.

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Mason, Ohio: Cengage Learning.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika,16,* 3, 297–334.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T., & MacNamara, T. (1999). *Dictionary of language testing*. Cambridge etc.: Cambridge University Press.

Ebel, R. L. (1954). Procedures for the Analysis of Classroom Tests. *Educational and Psychological Measurement*, no. 14, pp. 352–364. DOI: 10.1177/001316445401400215.

Hughes, A. (2003). *Testing for Langure Teachers*. Cambridge: Cambridge University Press.

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

Kostelecká, Y. & Jančařík, A. (2014). The Process of Czech Language Acquisition by Foreign Pupils at Lower Secondary School. *Journal on Efficiency and Responsibility in Education and Science*. ***7*** (1), pp. 7–13.

Kostelecká, Y. et al. (2013). *Žáci – cizinci v základních školách: Fakta, analýzy, diagnostika*. Praha: Univerzita Karlova v Praze, Pedagogická fakulta.

Kreidl, M. (2004). Metody měření reliability a validity. *Socioweb*. Retrieved from: http://www.socioweb.cz/index.php?disp=teorie&shw=153&lst=103

Kuder, G. F. & Richardson, M. W. (1937). The Theory of the Estimation of Test Reliability. *Psychometrika*. *3*(2), pp. 151–160. DOI: 10.1007/BF02288391. Retrieved from: http://link.springer.com/10.1007/BF02288391.

Lord, F. M., Novick, M. R., & Birnbaum, A. (2008). *Statistical theories of mental test scores*. Charlotte, N.C.: Information Age Pub.

Perclová, R. (2010). *Evropské jazykové portfolio, pro žáky a žákyně ve věku 11–15 let v České republice*. Praha: Fortuna Print.

McKay, P. (2008). *Assessing Young Language Learners*. Cambridge: Cambridge University Press.

Řehák, J. (1998). Quality of Data I. Classical Model of Measuring Reliability and Its Practical Application. *Sociologický časopis*. 34, pp. 51–60.

Soukup, P. Čím větší, tím lepší (aneb mýty o reliabilitě). *Socioweb*. Retrieved from: http://www.socioweb.cz/index.php?disp=teorie&shw=242&lst=108

Vacková, J. et al. (2002). *Evropské jazykové portfolio: pro žáky do 11 let v České republice*. Plzeň: Fraus.

Contact:
Mgr. Kateřina Vodičková, MA, Ph.D.
Univerzita Karlova v Praze
Ústav jazykové a odborné přípravy
Vratislavova 29/10
1280 00 Praha
E-mail: katerina.vodickova@ujop.cuni.cz

Mgr. Yvona Kostelecká, Ph.D.
Univerzita Karlova v Praze
Pedagogická fakulta
ÚPRPŠ
M. Rettigové 8
116 39 Praha 1
E-mail: yvona.kostelecka@pedf.cuni.cz