

Test Assessment: Grading Practice (A Comparison Study of Preservice English Teachers and Inservice English Teachers in the Czech Republic)

Helena Zitková¹, Jaroslav Myslivec²

The Czech Republic, University of Pardubice, Faculty of Arts and Philosophy

Abstract

The present study aims to explore the realm of test assessment by examining variations in grading by distinct evaluators within the context of the Czech Republic. Specifically, this research investigates both preservice and inservice teachers of English, comparing their grading practices. We conducted a quantitative data analysis in our survey. The data were collected through a questionnaire, where the participants were asked to assess a sample of the same completed progress test. To statistically compare the samples, we employed a non-parametric Mann-Whitney test, Levene's test and the χ^2 test, and a simple linear regression model. The research findings offer valuable insights into the subjectivity and variability of grades. However, the main results indicate no significant differences in terms of variability in scores or the grades assigned by preservice and inservice teachers to an identical test. This study is anticipated to contribute to the understanding of grading in/consistency within the context of English test assessment. Moreover, the outcomes may have implications for teacher training programs, curriculum development, and the enhancement of assessment practices. The study provides a basis for re-evaluating the predominant use of grades in Czech primary schools and emphasizes the need to incorporate this knowledge into preservice teacher education.

Keywords: test assessment; grading variability; preservice English teachers; inservice English teachers; evaluation subjectivity

1. Introduction

Assessment¹ is an integral part of any learning, making it a vital component of the educational process in schools. The primary purpose of school assessment is to accompany students on their journey of discovery and support them in their learning process as they move towards their goals. Consequently, there are noticeable tendencies towards prioritizing the formative role of assessment. However, it is noteworthy that summative assessment, involving the use of grades, is still prevalent in schools.

The assessment in the form of grades² in educational settings has been a topic of expert discussion for many years, encompassing areas such as reliability of grades, students' perceptions, teachers' beliefs about grades, standard-based grading, or the relationships between grades and tested achievement (Brookhart et al., 2016). Our focus lies in the realm of subjectivity and grading variability, aiming to illuminate potential differences in grading practices between preservice³ and inservice teachers of English.

1.1 Assessment in educational settings

Contemporary approaches to learning in educational settings demonstrate a noticeable shift from traditional evaluation of learning based on grades towards prioritizing assessment for learning, also known as formative assessment. It is widely believed that assessment should promote learning and enhance students' performance and achievements by providing valuable, high-quality feedback on their learning processes and outcomes in relation to given goals (Slavík, 2003; Wormeli, 2006; Volante & Fazio, 2007; Anderson, 2008; OECD, 2012; Žlábková & Rokos, 2013; William & Leahy, 2015; Crockett & Churches, 2017).

In essence, assessment serves as a key tool to provide valid and valuable information that helps students improve and support their learning processes. It influences students' motivation, attitudes towards school and school subjects, shapes student's self-concept and ambitions, has an impact on the teacher-student relationship, and can determine a social position in the classroom (Butler & Nisan, 1986; Slavík, 2003, p. 5; OECD, 2012). For these reasons, it should be carried out with an awareness of these effects.

Interestingly, grades, as a commonly used form of assessment in schools, do not solidly reflect learning and do not appear to offer valuable feedback to students. They merely indicate the

¹ In the context of education, some authors distinguish between the terms “evaluation” and “assessment” (Astin and Antonio, 2012), however, these terms will be used interchangeably in correspondence with, for example, Ur's (1991) or Harmer's (2007) notion within the scope of this paper.

² To prevent misunderstanding, the term “grade”, and not “mark”, will be used to refer to grades (grade ‘1’ through ‘5’) throughout the whole paper.

³ The terms preservice teachers and student teachers will be used interchangeably in the present paper.

level of subject matter mastery and often fail to deliver reliable information about students' progress in learning, while also suppressing creativity and diminishing interest (Kohn, 2011; Schinske & Tanner, 2014). Furthermore, research has confirmed that grades undermine intrinsic motivation and enhance anxiety, especially among struggling students, who experience repeated failure (Butler & Nisan, 1986; Kohn, 1999; Schinske & Tanner, 2014). As fittingly mentioned by Crockett and Churches (2017, p. 2-3) "a number... means little and accomplishes even less". In summary, grades have shortcomings in providing meaningful feedback, enhancing intrinsic motivation, and fostering progress in learning.

In line with modern approaches to assessment (Kohn, 2011; OECD, 2012; William and Leahy, 2015; Crockett & Churches, 2017), the core Czech curricular document for primary education (*Framework Educational Programme for Basic Education*) calls for "changes in the assessment of the students towards diagnostics on an ongoing basis, [...] and a wider use of verbal assessment", and demand assessment built on "evaluation of the student's individual development and on the capacity to make finely-tuned, positive evaluative judgements" (Ministry of Education, 2023, p. 6, 8). Despite these requirements and the acknowledgement that grades do not communicate clear and useful information to promote student learning, they are still widely used. As indicated in the Czech School Inspection Annual Report, the predominance of grading-focused assessment, emphasizing error identification over fostering active learning, continues to persist in Czech primary schools (Česká školní inspekce, 2022). The fact that assessment for learning in a form of valuable feedback is not often used in Czech schools, was also previously confirmed by OECD investigation (Santiago et al., 2012).

1.2 School assessment in Czech law and core curricular documents

To understand the entire context of assessment in Czech schools, the curricular and legislative background must be considered. School assessment is regulated by law, with the only legally binding forms being the mid-term and end-of-year report cards, where students' performance and achievement can be summatively assessed either using grades on a scale from 1 (the best) to 5 (the worst), or using written verbal feedback, or a combination of both (School Act, 2004; Decree, 2005). Currently, final grades on the school report card are one of the prerequisites for admission to higher education institutions, including secondary schools and some universities. Consequently, schools relying solely on verbal assessment must convert it into grades for higher education admission purposes (School Act, 2004). As such, grades play a crucial role in the Czech school system, significantly influencing the future trajectory of students' lives.

By law, teachers are expected to assess students' performance and achievements in alignment with expected outcomes specified in their *School Educational Programme*, which must correspond to the core curriculum document (*Framework Educational Programme for Basic Education*). Rules and criteria for assessment at each grade level are mandatory components of every school's guidelines. However, due to the need to accommodate the diverse content across various school subjects, the formulation of these criteria is so general and barely

measurable that it is not possible for all teachers to perceive them in the same way. This generality and lack of precise quantification are illustrated in the following example, presenting formulated assessment criteria for the '2' grade level in subjects with a predominantly theoretical focus, such as mathematics, geography, history, or languages:

The student demonstrates comprehensive control over the required knowledge, facts, concepts, and definitions. The student works independently and productively when applying acquired knowledge and skills to solve both theoretical and practical tasks, either autonomously or with minor prompts and guidance from the teacher. The student thinks correctly; their thinking shows logic and creativity, occasionally originality. Oral and written communication tends to have minor imperfections in terms of correctness, accuracy, and conciseness. The quality of the student's performance is usually without significant mistakes. (Classification set of criteria established by a randomly selected primary school in the Czech Republic)

These criteria can be interpreted from different perspectives, allowing every teacher to evaluate promptness, productivity, and the significance of mistakes in their own unique way. Likewise, judgments about correctness, creativity, logic, and originality in thinking can vary from teacher to teacher. Moreover, everyone may have a different concept of the necessary level of help and guidance.

Furthermore, the legislation does not specify either specific assessment methods or prescribed assessment scales. Thus, in the absence of standardized exams or tests, which is atypical in daily school practice, sets of criteria, distribution of points in assessment scales, and grading scales can be perceived and operated diversely (Slavík, 2003). Therefore, it is common for teachers to create their own set of criteria and assessment scales within the boundaries of the school's general guidelines. While teachers have the autonomy to manipulate these criteria freely, it results in low objectivity in the assessment (Žlábková & Rokos, 2013). This practice falls within their full competence and autonomy. Yet, the strategies employed by Czech teachers in developing their sets of criteria and assessment scales have not been thoroughly investigated, leaving the current knowledge of this issue insufficient.

1.3 Subjectivity of assessment

In the history of school assessment research, the subjectivity of grading has already been indicated (Starch & Elliot, 1912; Brimi, 2011), with various aspects potentially influencing the grading process and providing misleading information, thus influencing the grade. Notably, there are substantial discrepancies in the criteria and procedures teachers employ when assessing learning outcomes (Meadows & Billington, 2005; Brookhart and Guskey in Guskey and Brookhart, 2019). Additionally, the cognitive process and reasoning behind grade assignments extend beyond conscious control, making it challenging for teachers to pinpoint the specific criteria influencing the grade (McMillan, 2003; Slavík, 2003; Svennberg et al., 2014). Consequently, grades may not solely reflect the level of achieved learning outcomes but may also encompass a combination of non-cognitive (also non-achievement) aspects,

representing diverse conscious and unconscious perspectives, beliefs, and values held by different teachers (Cox, 2011; Brookhart et al., 2016). Researchers exploring grading practices have identified several non-cognitive factors impacting grades, such as behavior and work habits, improvement, motivation and engagement, or personal qualities of a student (Wormeli, 2006; Duncan & Noonan, 2007; Schinske & Tanner, 2014; Svennberg et al., 2014). The perception of these factors by teachers is just a matter of the individual and personal point of view, rendering assessment intuitive and subjective (Slavík, 2003; Hay & McDonald, 2008).

In addition, grades can be influenced by factors like the number of tests to be corrected, teacher mental fatigue, or the length of teaching experience and teacher training (Klein, 2002; Meadows & Billington, 2005). Studies suggest that inexperienced assessors tend to grade more severely than experienced ones, as pointed out by Meadows and Billington (2005). Weigle (1999) found that when assessing a writing task, inexperienced raters were more severe compared to experienced raters. Also, inservice teachers may perceive themselves as more skillful in using performance assessments and grading than student teachers (Alkharusi, 2011), potentially leading to differences in grading practices. Moreover, despite teacher training in using the same grading methods, inconsistencies in grades may persist (Brimi, 2011). On the other hand, Mertler' survey (2004) found no significant differences between inservice and preservice teachers in developing valid grading procedures.

The combination of all the mentioned factors complicates the interpretation of grades not only across schools but also within the same school. Thus, grades are not a comparable and objective measure of the knowledge and skills attained by different students, given the wide variations in assessment rules and practices from school to school and teacher to teacher (Duncan & Noonan, 2007; OECD, 2012; Žlábková & Rokos, 2013; Brookhart et al. 2016; Federičová, 2019). In our study, we investigated the grading practices of student teachers and inservice teachers from different schools.

2. Methods

The study aimed to examine grading practices and the variability of grades between preservice and inservice teachers. Specifically, we focused on the assessment criteria and grading scales that preservice and inservice teachers applied when assessing the same non-standardized progress test. While inservice teachers are expected to be professionals with expertise and experience, research in the field of assessment practices has indicated differences in grading practices between preservice and inservice teachers. Building on existing research (Meadows & Billington, 2005; Weigle, 1999), we anticipate that student teachers may evaluate the test more severely than inservice teachers. We also predict that student teachers will exhibit greater variability in test scores compared to inservice teachers, reflecting the differences in their grading practices and experience levels. The research employs a quantitative analysis of assessment criteria and grading scales gathered through a questionnaire.

2.1 Survey participants

The preservice teachers surveyed were in their second year of a bachelor's study program *English for Education*, preparing them to become English teachers⁴. None of them had any school teaching experience, and thus, they had no direct practice with evaluating and grading tests.

A total of 30 students enrolled in the ELT (English Language Teaching) methodology course in the summer semester of 2020 were given a survey questionnaire. Out of the 30 questionnaires submitted, only 19 could be used due to incomplete data.

Inservice lower-secondary teachers of English were surveyed in 2021. The survey questionnaire was distributed online via social media, specifically targeting English teachers. A total of 32 English teachers responded to and completed the questionnaire, but 7 respectively 8 respondents⁵ were discarded, either due to incomplete data or because they did not meet the criteria for teaching at a lower-secondary school.

As a result, the study worked with two survey sample: Survey sample I (N=19) representing preservice teachers and Survey sample II (N=25) representing inservice teachers (with one grading scale with the number of points and one grade missing).

All respondents were of Czech origin. Inservice teachers were asked for background information, including details about the length of their teaching experience, English language teacher qualifications, and whether they teach English at the lower-secondary school level. Additional information, such as gender and age of all the respondents, as well as the place of work and educational background for inservice teachers, was considered irrelevant to this study and therefore was not included in the survey.

2.2 Data Collection

All participants received a questionnaire that included a sample of a completed progress test⁶ and questions in which they were asked to assess it. Specifically, respondents were instructed to:

⁴ The data applicable for the present study was selected from a broader qualitative study conducted in 2020, which aimed to examine the actual perception of pre-service teachers regarding assessment and grading (see Žitková, H. (2021) "Assessment and Grading - Perception of Pre-Service English Teachers". *Hradec Králové Journal of Anglophone Studies*. Univerzita Hradec Králové. 8(1-2), pp. 31-55. Available: http://pdf.uhk.cz/hkjas/pi/pdf/vol8nr1_2021.pdf#page=31)

⁵ One respondent provided only a grading scale in percentage without indicating the number of points as well as the grade. Therefore, we used this respondent's data only when analyzing the grading scales and excluded it from the analysis of grades.

⁶ The sample of a completed progress test (inspired by Ur, 1991) is appended to the text.

1. create a set of criteria for assessing the test
2. determine the number of points for fulfilling each criterion
3. design a grading scale
4. assess the test and assign a grade ranging from 1 to 5.

Participants were provided with necessary information, indicating that the test was completed by a ninth grader, approximately at A2 level, and aimed to evaluate the understanding of specific words' meaning and the correct usage of relative pronouns when defining the words.

Some attempts to describe and explain the decision-making process behind creating the criteria and determining the points were indicated in the answers, but not all the respondents included them. Therefore, this data was deemed irrelevant to the study and was omitted from the analysis.

2.3 Results

A quantitative data analysis was carried out in the survey. First, the results of a descriptive analysis of the data collected separately from preservice teachers (Survey sample I) and inservice teachers (Survey sample II) will be presented. Subsequently, the data obtained from all three survey samples will be statistically compared.

2.3.1 Display of grading scales - preservice teachers

From Survey sample I (N=19), a total of 12 different grading scales were identified. They are all presented in three tables below based on the maximum possible points. Table 1 contains grading scales with 6 maximum points, Table 2 displays grading scales with 12 maximum points, and Table 3 showcases grading scales with 15 and 18 maximum points.

The most common grading scale was designed by 6 student teachers, marked gray in Table 1. Grading scales marked gray in Table 2 were each represented twice. Grading scales marked white in all tables were represented only once.

In addition, not only do the grading scales differ in the number of maximum possible points that can be given to the test, but they also vary in the distribution of points to grades from 1 to 5. In the following tables, all the grading scales designed by preservice teachers and corresponding grades are presented:

Table 1: Grading scales with maximum points of 6 and corresponding grades

Grade	Grading Scales – 6 points		
1	6	6-5.5	6-5
2	5.5-5	5-4	4.5 – 3.5
3	4.5-4	3.5-2	3-2.5
4	3.5-3	1.5-0.5	2-1
5	2.5-0	0	0.5 - 0

Table 1 illustrates the variations in point distribution within grading scales with a maximum of 6 points. As shown, the same test can receive different grades depending on the grading scale applied. For instance, a test earning 2 points could receive three distinct grades (5, 3, and 4) from three different assessors. This implies that a learner could receive three different grades from different teachers for the exact same performance.

Table 2: Grading scales with maximum points of 12 and corresponding grades

Grade	Grading Scales – 12 points					
1	12	12	12-11	12-10	12-10	12-10.5
2	11	11-10	10-9	9-7	9.5-8.5	10-8.5
3	10	9-8	8-7	6-4	8-7	8-6.5
4	9	7-6	6-5	3-2	6.5-5.5	6-4
5	8-0	5-0	4-0	1-0	5-0	3.5-0

As evident from Table 2, the same principle of assigning different grades to an identical test is applicable within grading scales with a maximum of 12 points. For example, a test scoring 8 points would be given three different grades (5, 3, and 2).

Table 3: Grading scales with maximum points of 15 and 18 and corresponding grades

Grade	Grading Scales – 15 points		Grading Scale – 18 points
1	15-14	15-12	18-16
2	13.5-10.5	11.5-9.5	15-12
3	10-8	9-6	11-8
4	7.5-5.5	5.5-2.5	7-4
5	5-0	2-0	3-0

Table 3 illustrates that when utilizing grading scales with a maximum of 15 points, the same test would not consistently receive the same grade. In this case, the difference in grades would always be just one. Compared to the grading scale having an 18-point maximum, it is noteworthy that a test with 8 points would consistently receive the same grade, regardless of whether the maximum possible points for the entire test were 15 or 18.

2.3.2 Display of grading scales - inservice teachers

From survey sample II (N=24), a total of 22 different grading scales emerged. They are presented in the three tables below based on the maximum possible points. Table 4 displays grading scales with maximum points of 18, 15 and 24. Table 5 presents grading scales with

maximum points of 12, while Table 6 includes two grading scales with a maximum of 6 points. Grading scales marked gray in Table 4 and Table 5 were each represented twice, while those marked white in all tables were represented just once.

Table 4: Grading scales with maximum points of 18, 15 and 24 and corresponding grades

Grade	Grading Scales – 18 points					15 pts.		24 pts.
1	18-17	18-16	18-16	18-16	18-16	15-13	15-13	24-20
2	16-13	15-13	15-13	15-13	15-13	12,5-10,5	12,5-11	19-15
3	12-8	12-9	12-9	12-9	11-9	10-7,5	10,5-6,5	14-10
4	7-4	9-5	8-6	8-4	8-4	7-5	6-4	9-5
5	3-0	4-0	5-0	3-0	3-0	4,5-0	4-0	4-0

Table 4 demonstrates that when using grading scales with a maximum score of 18 points, the same test would not consistently receive the same grade, as indicated by variations in grades for 16, 8, 5, and 4 points awarded. In this case, the difference in grades would always be just one. The same goes for the grading scale having a 15-point maximum. It is noticeable that a test with 5 points would consistently receive the same grade, regardless of whether the maximum possible points for the entire test were 18, 15, or 24.

Table 5: Grading scales with maximum points of 12 and corresponding grades

Grade	Grading Scales – 12 points					
1	12-10,5	12-10,5	12-10,5	12-10,5	12-11	12-11
2	10-9	10-8	10-9	10-8,5	10-9	10-9
3	8,5-7,5	7,5-5,5	8-6	8-4,5	8-5	8-7
4	7-6	5-3	5-3	4-2,5	4-3	6-5
5	5,5-0	2,5-0	2-0	2-0	2-0	4-0
1	12-11	12-11	12-11	12-10	12-10	12-10
2	10-9	10-9	10-8	9-8	9-8	9-8
3	8-6	8-6	7-5	7-6	7-6	7-5
4	5-4	5-3	4-2	5-3	5-4	4-3
5	3-0	2-0	1-0	2-0	3-0	2-0

As Table 5 illustrates, the variations in point distribution within grading scales with a maximum of 12 points are diverse. The same test would exhibit grade inconsistencies, as shown by variations in grades assigned for scores of 16, 8, 5, and 4 points. Moreover, as depicted, depending on the scale used, the same test can receive different grades. For example, a test achieving a score of 5,5 points could be assigned three distinct grades (3, 4, and 5). This suggests that a student could receive three different grades from different teachers for the very same performance.

Table 6: Grading scales with maximum points of 6 and corresponding grades

Grade	Grading Scales – 6 points
1	6-5
2	4
3	3
4	2
5	1
1	6-5,5
1-	5
2	4,5
2-	4
3	3,5
3-	3
4	2,5
4-	2
5	1,5-0

Note: The assessment, in which a minus is used after the grade, is not official according to Czech legislation.

As shown in Table 6, the same principle of assigning different grades to an identical test is applicable even within two grading scales with a maximum of 6 points. However, this finding is distorted by the grading scale with minuses.

To summarize the findings, when comparing the numbers of identified grading scales and score ranges between survey samples, inservice teachers appear to exhibit less consistency in designing grading scales. In contrast to those created by preservice teachers, the grading scales and score ranges produced by inservice teachers demonstrate greater variability.

2.3.3 Comparison between preservice and inservice teachers

Since the data did not show normal distribution (verified by Shapiro-Wilk test, $p < 0.0001$), a non-parametric Mann-Whitney test was used to compare the samples. Levene's test was used to compare the variability of the samples. For categorical data, the χ^2 test was used. A simple linear regression model was also used.

Percentage of points and grades

Starting with the points scored by the respondents in the test, the percentage of minimum and maximum scores, and the corresponding grades were calculated for all three survey samples. Table 7 summarizes descriptive statistics for inservice and preservice teachers' assessment of the test.

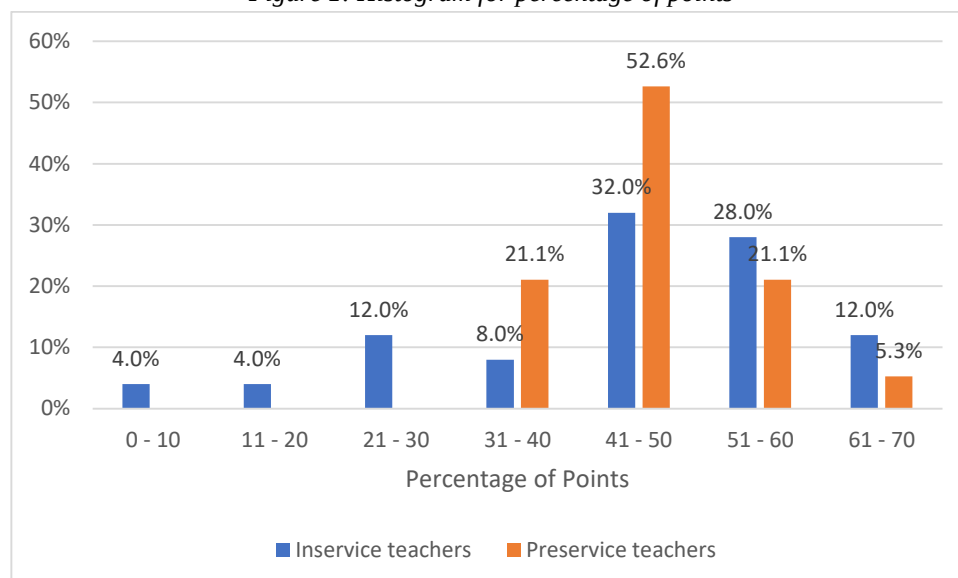
Table 7: Descriptive statistics

	Inservice teachers		Preservice teachers	
	Percentage of points	Grade	Percentage of points	Grade
N	25	24	19	19
min	0	2	33.00	2
max	66.67	5	67.00	5
M	45.03	3.43	47.74	3.63
median	50	3	50	4
SD	16.73	0.83	9.05	0.96
CV	0.37	0.24	0.19	0.26

Note: M = mean; SD = standard deviation; CV = coefficient of variation

Figure 1 and 2 show histograms for percentage of points and grades.

Figure 1: Histogram for percentage of points

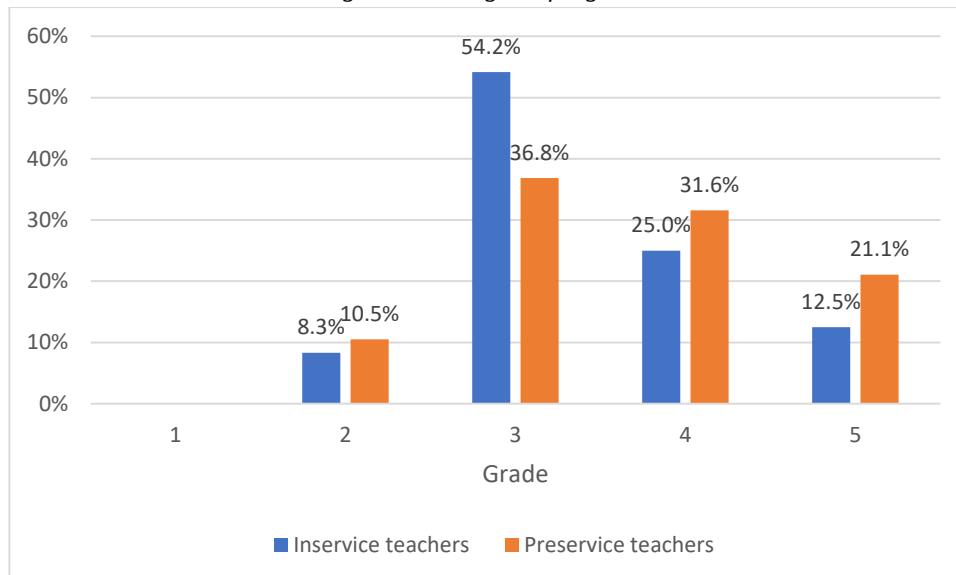


In the case of preservice teachers, the percentage of points ranges from 0% to 66.67%, with a mean of 45.03% and a median of 50%. The sample standard deviation is 16.73. 60% of respondents scored between 41% and 60%, 12% scored more than 60%, and 28% scored 40% or less (see Figure 1). A respondent who gave 0% (outside 1.5 times the interquartile range below the lower quartile) can be considered an outlier.

As for the inservice teachers, the percentage of points ranges from 33% to 67%, with a mean of 47.74%, a median of 50%, and a sample standard deviation of 9.05 (Table 7). 73.7% of respondents gave 41 to 60% points, 5.3% gave more than 60%, and 21.1% gave 40% points or less (see Figure 1).

Figure 2 shows the percentages of teachers and students with the appropriate grade they awarded the test.

Figure 2: Histogram for grades



Note: One grade of 3 minus was included among the 3

Grades for inservice teachers range from 2 to 5, with a mean grade of 3.44, a median of 3, and a sample standard deviation of 0.83 (see Table 7). 54.2% of respondents gave a grade of 3, and over 79% of respondents gave a 3 or 4 (see Figure 2).

For preservice teachers, grades also range from 2 to 5, with a mean grade of 3.63, a median of 4, and a sample standard deviation of 0.96 (see Table 7). 36.8% of respondents gave a grade of 3, and 68.4% of respondents gave a grade of 3 or 4 (see Figure 2).

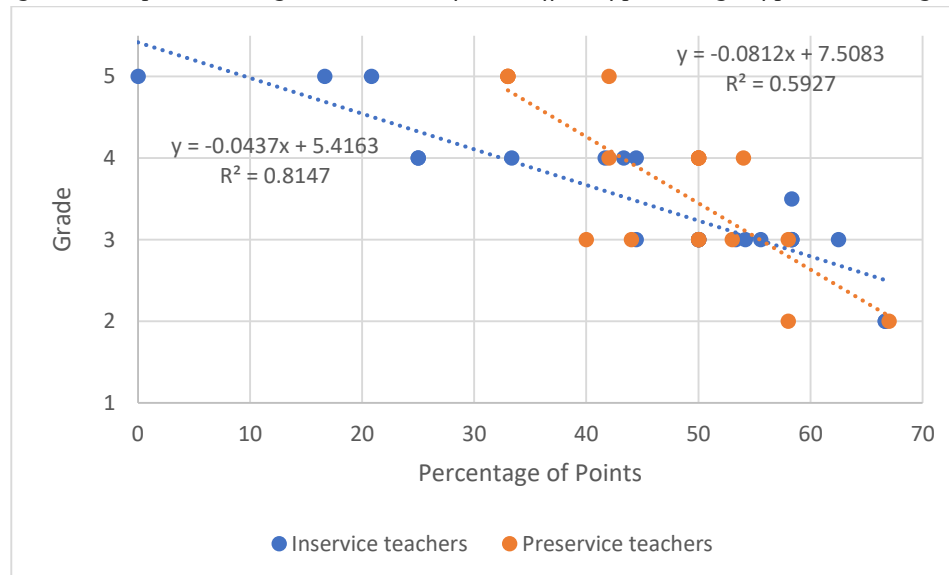
Although Figure 2 might suggest that preservice teachers rate the test more severely than inservice teachers and that there is greater variability in the scores given by preservice teachers, these differences are not statistically significant.

In summary, the differences in scores are not significant in either percentage of points ($p = 0.867$) or grades ($p = 0.475$). Inservice teachers and preservice teachers do not differ either in terms of variability in percentage points ($p = 0.064$) or in terms of grades ($p = 0.376$). Thus, neither the hypothesis of greater preservice teacher variability in test scores nor the hypothesis of more severe test scores by preservice teachers was supported.

Interestingly, the variability in scores is slightly higher for inservice teachers, whereas the variability in grades is slightly higher for preservice teachers (although the differences are not significant). This led us to the idea of comparing inservice and preservice teachers also in terms of how percentage points are converted into grades. We therefore constructed a simple linear

regression model for both groups of respondents that shows the relationship between percentage points and the resulting grade (see Figure 3).

Figure 3: Simple linear regression models for the effect of percentage of points on the grade



Inservice teachers are much more consistent in assigning grades based on percentage points. While 81.5% of the variability in grades for inservice teachers can be attributed to variability in percentage points, for preservice teachers it is only 59.3%.

Table 8 shows the grades awarded by inservice teachers and preservice teachers for given percentages.

Table 8: Percentages of points and grades

Grade	Percentage of Points	
	Inservice teachers	Preservice teachers
2	66.67	58–67
3	44.44–62.5	40–58
4	25–44.44	42–54
5	0–20.83	33–42

Inservice teachers consistently assigned grades using percentage points, with two exceptions (see Figure 3 and Table 8). The first exception is a respondent who assigned a three-minus grade for 58.33 percent of the points, while others awarded a three for anywhere between 44.44 to 62.5 percent of the points. The second exception involves two respondents, one of whom assigned a three for 44.44 percent of the points, while the other gave a four.

For preservice teachers, the overlap of the intervals is substantially higher (see Figure 3 and Table 8). The entire percentage point interval for which some respondents gave a four (42-54 percentage points) is a subset of the interval for which other respondents assigned a three (40-

58 percentage points). Moreover, both intervals partially overlap with the interval for a five (33-42 percentage points). Thus, for example, 42 percentage points could receive a three, a four, or a five. In addition, for 58 percentage points, one respondent awarded a three, and another assigned a two.

Good grades and bad grades

We categorized the grades into two groups: 'good grades' (2 or 3) and 'bad grades' (4 or 5) and attempted to determine whether there existed a difference in the test scores between inservice and preservice teachers using this categorization. The results are presented in Table 9.

Table 9: Numbers of inservice and preservice teachers and test scores

	Good grade	Bad grade	Suma
Inservice teachers	15 62.50 %	9 37.50 %	24
Preservice teachers	9 47.37 %	10 52.63 %	19
Suma	24	19	43

Although inservice teachers rate the test more by 'good grades' and preservice teachers more by 'bad grades', still the differences are not statistically significant ($p = 0.321$).

2.4 Discussion

Based on our research findings, it is evident that the same test can receive different scores from different teachers. Research in the Czech Republic has also shown significant differences in grading practices within primary schools (Santiago et al., 2012; Münich and Protivínský, 2018). On average, there can be a noticeable distinction of up to an entire final grade assigned between different schools (Münich & Protivínský, 2022). In our survey samples, a single test yielded grades ranging from 2 to 5, emerging from distinct grading scales designed by both preservice and inservice teachers. From a summative assessment perspective, this could result in substantial discrepancies in the final grades awarded by teachers. Such grading subjectivity can greatly impact a student's future academic path (Slavík, 2003; Hay & McDonald, 2008) as aptly noted by Starch and Elliot (1912) in their research conclusions, where they highlighted that a student's progress depends on their teacher's subjective estimate.

Until today, studies have investigated grading reliability (Starch & Elliot, 1912; Duncan & Noonan, 2007; Brimi 2011) and compared preservice and inservice teachers' assessment literacy (Mertler, 2004). Nevertheless, we see a research gap in examining the reliability of grading the same progress test in classes of English as a foreign language by non-native

teachers, rather than grading the same student paper by native English teachers in an anglophone environment, as in the aforementioned studies.

It is important to note that due to the limited number of respondents, the findings of this study cannot be generalized to other groups of inservice and preservice teachers. Additionally, the available data impose limitations on the results. For instance, our analyses do not, and cannot, purport to identify the various aspects of the curriculum that teachers assessed, or the corresponding weights teachers assign to them. Furthermore, we cannot correlate differences in grades with teacher characteristics. Moreover, it remains unclear on what specific criteria the teachers based their grading. While some mentioned, 'grammar', they did not specify whether they meant, for example, spelling, word order, or the use of articles. Finally, for the purpose of our research, we adapted a test prepared by Penny Ur (1991) without knowing its true origin and background information. This lack of knowledge may also mark a weakness of the study.

As new paths for studies in this area, we suggest replicating the study with a larger sample, which would enhance the validity of results. The next study could also benefit from utilizing a real test written by a real learner. Another course of action will be to focus on uniformity of criteria. In our study, respondents were not given criteria by which to rate the text. It would be interesting to explore what the variability in ratings will be if these criteria were provided to respondents.

Furthermore, we propose several additional research directions. Systematic research investigating the factors influencing grading reliability, as suggested by Meadows and Billington (2005), could yield valuable insights. Also, it would be valuable to explore what teachers of English include in the final grade and perhaps also delve into teachers' internal beliefs and values, and their decision-making rationale for using assessment practices, as suggested by Duncan and Nunan (2017).

3. Conclusion

By conducting a detailed examination of how preservice and inservice English teachers approach the assessment of an identical test, this study provides valuable insights into the subjective nature and variations in test evaluation. While the results may be problematic in terms of generalizing and may not reach statistical significance, they underscore the subjectivity, variability, and implications of grading. Despite the lack of statistical significance, it remains crucial for individual students whether they receive a grade of 1, 2, 3, 4 or 5 on a test. It is obvious that a student can get different grades from different teachers for the same performance, and thus, teachers can influence students' lives with their subjective judgements, which can shape their future life.

We firmly believe that enhancing assessment literacy is essential for both preservice and inservice teachers. Our research can serve as a valuable starting point for initiating discussion about assessment in teacher training programs and help them to become confident in making informed assessment decisions. Moreover, it may prompt reflection on the meaningfulness of grades in schools. While some authors, such as Kohn (2011) advocate for the complete abolition of grades, the OECD (2012) recommends a gradual phase-out of grades. Until their complete elimination, it suggests implementing collaborative grading approaches involving students.

In conclusion, although the results of the research do not show significant differences, this research nonetheless presents valuable insights into the subjectivity and variability of grades and thus offers space for re-evaluating the predominant use of grades in Czech primary schools and the need to incorporate this knowledge into preservice teacher education.

References

- Alkharusi, H. (2011). "Self-Perceived Assessment Skills of Pre-service and In-service Teachers," *Jurnal Pendidikan Malaysia*, vol. 36(2), pp. 9-17.
- Anderson V. J. (2008). "Grading." In *Encyclopedia of Educational Psychology*. Thousand Oaks, CA: Sage; pp. 450-453. [Online]. Available: https://scholar.google.com/scholar_lookup?title=In:+Encyclopedia+of+Educational+Psychology&author=VJ+Anderson&publication_year=2008&
- Astin, A.W. and Antonio, A. L. (2012). *Assessment for Excellence: The Philosophy and Practice of Assessment and Evaluation in Higher Education*. American Council of Education and Rowman and Littlefield Publishers, Inc.
- Brimi, H. M. (2011). "Reliability of Grading High School Work in English," *Practical Assessment, Research, and Evaluation*, vol. 16(17), pp. 1–12. [Online]. Available: <https://scholarworks.umass.edu/pare/vol16/iss1/17>
- Brookhart, S. M., Guskey, T. R., Bowers, A. J., McMillan, J. H., Smith, J. K., Smith, L. F., Stevens, M. T. and Welsh, M. E. (2016). "A Century of Grading Research: Meaning and Value in the Most Common Educational Measure," *Review of Educational Research*, 86(4), pp. 803–848. [Online]. Available: <https://doi.org/10.3102/0034654316672069>
- Brookhart, S. M. and Guskey, T. R. (2019). "Reliability in Grading and Grading Scales". In *What We Know About Grading. What Works, What Doesn't, and What's Next*. edited by Guskey, T. R. and Brookhart, S.M. Alexandria, VA: ASCD.
- Butler, R. and Nisan, M. (1986). "Effects of no feedback, task-related comments, and grades on intrinsic motivation and performance," *Journal of Educational Psychology*, vol. 78(3), pp. 210–216.
- Cox, K. B. (2011). "Putting classroom grading on the table, a reform in progress," *American Secondary Education*, vol. 40(1), pp. 67–87. [Online]. Available: https://hhs.sau70.org/uploaded/Hanover_High_School/Library/Staff_Research/Putting_Classroom_Grading_on_the_Table.pdf
- Crockett, W. L. and Churches, A. (2017). *Mindful Assessment. The 6 Essential Fluencies of Innovative Learning*. Solution Tree Press.
- Česká školní inspekce (ČŠI). (2022). *Kvalita vzdělávání v České republice 2021/22. Výroční zpráva*. Praha. Available: https://www.csicr.cz/CSICR/media/Prilohy/2022_p%c5%99%c3%adlohy/Dokumenty/Vyrocn%C3%AD-zprava-2021-2022-everze.pdf

Decree No 48/2005 Sb. - Vyhláška o základním vzdělávání a některých náležitostech plnění povinné školní docházky. (2005). *Sbírka zákonů*, částka 11/2005.

Duncan, R. C. and Noonan, B. (2007). "Factors affecting teachers' grading and assessment practices," *Alberta Journal of Educational Research*, vol. 53, pp. 1–21.

Federičová, M. (2019). *Dopady známek na vysvědčení na životní rozhodnutí žáků*. IDEA 1/2019.

Hay, P. J. and Macdonald, D. (2008). "(Mis)appropriations of criteria and standards-referenced assessment in a performance-based subject," *Assessment in Education: Principles, Policy & Practice*, vol. 15, pp. 153–168.

Harmer, J. (2007). *The Practice of English Language Teaching*. Pearson Longman.

Klein, J. (2002). "The failure of a decision support system: inconsistency in test grading by teachers," *Teaching and Teacher Education*, vol. 18(8), pp. 1023-1033.

Kohn, A. (1999). *Punished by rewards: The trouble with gold stars, incentive plans, As, and other bribes*. Boston: Houghton Mifflin.

Kohn, A. (November 2011). "The Case against Grades," *Educational Leadership*. [Online]. Available: <https://www.alfiekohn.org/article/case-grades/>

Marzano, R. J. (2000). *Transforming Classroom Grading*. ASCD, Alexandria, VA, 1.

McMillan, J. H. (2003). "Understanding and improving teachers' classroom assessment decision-making: Implications for theory and practice," *Educational Measurement: Issues and Practice*, 22(4), pp. 34-43.

Meadows M. and Billington L. (2005). *A Review of the Literature on Marking Reliability*. National Assessment Agency. [Online]. Available: https://filestore.aqa.org.uk/content/research/CERP_RP_MM_01052005.pdf

Mertler, C. A. (2004). "Secondary Teachers' Assessment Literacy: Does Classroom Experience Make A Difference?," *American Secondary Education*, vol. 33(1), pp. 49–64. [Online]. Available: <http://www.jstor.org/stable/41064623>

Ministry of Education. (2023). *Framework Educational Programme for Basic Education (FEP BE)*. Praha.

Münich, D. and Protivínský, T. (2018). *Co skrývají známky na vysvědčení?* IDEA.

Münich, D. and Protivínský, T. (2022). *Rozdíly v přísnosti známkování žáků a dopady na vzdělanostní aspirace*. Národohospodářský ústav AV ČR.

OECD (2012). *Grade Expectations: How Marks and Education Policies Shape Students' Ambitions*. [Online]. PISA, OECD Publishing. Available: https://read.oecd-ilibrary.org/education/grade-expectations_9789264187528-en#page58

Santiago, P., Gilmore, A., Nusche, D. and Sammons, P. (2012). *OECD Reviews of Evaluation and Assessment in Education: Czech Republic 2012*. [Online]. OECD Publishing. Available: www.oecd.org/dataoecd/33/47/49479976.pdf

Schinske J. and Tanner K. (2014). "Teaching More by Grading Less (or Differently)," *CBE-Life Science Education*, vol. 13(2), pp.159-66. [Online]. Available: <https://www.lifescied.org/doi/epdf/10.1187/cbe.cbe-14-03-0054>

School Act No 561/2004 Sb. - Zákon o předškolním, základním, středním, vyšším odborném a jiném vzdělávání - školský zákon. (2004). *Sbírka zákonů*, částka 190/2004.

Slavík, J. (2003). "Autonomní a heteronomní pojetí školního hodnocení - Aktuální problém pedagogické teorie a praxe," *Pedagogika*, vol. 53(1), pp. 5-25.

Starch, D. and Elliot, E. (1912). "Reliability in grading high school work in English," *School Review*, vol. 20, pp. 442-457.

Svennberg, L., Meckbach, J. and Redelius, K. (2014). "Exploring PE teachers' 'gut feelings': An attempt to verbalise and discuss teachers' internalised grading criteria," *European Physical Education Review*, vol. 20, pp. 199–214.

Ur, P. (1991). *A Course in Language Teaching. Practice and Theory*. CUP.

Volante, L. and Fazio, X. (2007). "Exploring Teacher Candidates' Assessment Literacy: Implications for Teacher Education Reform and Professional Development," *Canadian Journal of Education*, vol. 30(3), pp. 749-770.

Weigle, S. C. (1999). "Investigating Rater/Prompt Interactions in Writing Assessment: Quantitative and Qualitative Approaches," *Assessing Writing*, vol. 6(2), pp. 145-178

William, D. and Leahy, S. (2015). *Embedding Formative Assessment: Practical Techniques for K-12 Classrooms*. Learning Sciences International.

Wormeli, R. (2006). "Accountability: Teaching through assessment and feedback, not grading," *American Secondary Education*, vol. 34(3), pp. 14-27.

Zitková, H. (2021) "Assessment and Grading - Perception of preservice English Teacher,"
Hradec Králové Journal of Anglophone Studies, vol. 8(1-2), pp. 31-55. [Online]. Available:
http://pdf.uhk.cz/hkjas/pi/pdf/vol8nr1_2021.pdf#page=31

Žlábková, I. and Rokos, L. (2013). "Pohledy na formativní a sumativní hodnocení žáka v
českých publikacích," *Pedagogika*, vol. 63(3), pp. 328-354.

Appendix: The sample of a completed progress test (inspired by Ur, 1991):

Test on vocabulary and relative clauses, written during the lesson.

Define the following words, using who / which / that / whose / when / where.

For example: a deserted house = a house where nobody lives

- 1 courage: a man who not have any fear
- 2 an illusion: a false sight.
- 3 sweat: its like terrible but more than this
- 4 a virus: a flunk which make people sick
- 5 a paw: a part of a animal
- 6 a temple: a house where religious people lives in