DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„The application of assessment scales for writing and implications for teaching and testing in the Austrian EFL classroom: fairness, objectivity, and rater bias“

verfasst von / submitted by

Monika Schuster

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Magistra der Philosophie (Mag. phil.)

Wien, 2019 / Vienna, 2019

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 190 344 313
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	UF Englisch, UF Geschichte, Sozialkunde und Politische Bildung
Betreut von / Supervisor:	Mag. Dr. Helen Heaney, BA

Danksagung

Die Menschen, die mich täglich in meinem Leben begleiten, beeinflussen dieses mehr als ich ihnen jemals klarmachen könnte. Sie zeigen mir jeden Tag wie viel Kraft in uns steckt und wie stark wir sein können. Ihre Unterstützung und Ermutigungen geben mir stets das Gefühl, dass nichts unmöglich ist und sie haben mich mein ganzes Studium hindurch begleitet.

Ich danke meiner Betreuerin, Mag. Dr. Helen Heaney, BA, deren interessante Lehrveranstaltungen und hilfreiches Feedback mein Leben als Lehrerin für immer beeinflussen werden. Durch sie habe ich ein erstrebenswertes Vorbild gefunden, das reich an Geduld, Genauigkeit und Professionalität ist und ihren Studenten Wärme, Verständnis und Menschlichkeit entgegenbringt.

Ich danke meinen Geschwistern, vor allem meiner Schwester, Marlene Schuster, ohne deren Verständnis, Aufopferung und Optimismus ich niemals durch mein Studium gekommen wäre. Du bist mein Vorbild, mein Anker, mein Herz und meine Seele. Ebenso danke ich meinem Bruder, Matthias Schuster, der immer zu mir hält und mir stets seine Tür offenhielt. Ohne dein Obdach wäre diese Arbeit niemals fertig geworden.

Ich danke meinem Vater, Erich Schuster, dessen bedingungslose Unterstützung und Liebe mich immer daran erinnert, dass er an mich glaubt. Jeder meiner Erfolge hat ihren Ursprung in deinem Einfluss auf mich.

Zu guter Letzt danke ich meinem Freund, Christian Denk, dessen enorme Geduld, Zuspruch und Liebe mich immer wieder aufs Neue Kraft schöpfen lassen und der mir selbst in den dunkelsten Momenten zeigt, dass wir gemeinsam alles schaffen können.

Table of contents

List of abbreviations

BIFIE	Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens
BMBWF	Bundesministerium für Bildung, Wissenschaft und Forschung
CEFR	Common European Framework of Reference
CLAAS	CEFR Linked Austrian Assessment Scale
LBVO	Leistungsbeurteilungsverordnung
LSA	Lexical and structural accuracy
LSR	Lexical and structural range
OL	Organisation and layout
TA	Task achievement

List of figures

1. Introduction

In Austria, the focus of teaching and testing changed in the past years due to the implementation of standardised tests, competence-oriented syllabi, and new standards of education in general. Therefore, the EFL (English as a foreign language) classroom is currently undergoing a process which is influencing almost every aspect of teaching, and especially testing, English. Introducing an analytic rating scale for scoring written compositions in English has changed the process of assessing texts for EFL teachers.

Before these changes occurred, raters often graded essays by relying on a general impression of the text or they frequently counted errors, which might have led to unfair or unreliable grading since every teacher has their own idea of what a good text should look like. The newly introduced analytic rating scale, divided into four main categories replaces general impression marking with an equal evaluation of the components of successful writing. Therefore, a positive feature or one error or missing aspect should only lead to a higher or lower band in one of the four categories, giving the student the chance to score highly in one criterion even if there are problems in other criteria or the other way around so that the strengths and weakness of the student can be determined. This scale is a necessary and useful tool in order to establish a fair and comprehensible system. However, teachers must be trained to work with it and learn how to put their general impression grading and personal preferences aside. This issue of teachers' personal preferences in the process of grading written compositions with the analytic rating scale and its consequences for teaching and assessing is the main interest of this research.

To obtain relevant information which can be used to analyse the connection between the rating process and the influence of the raters' personal preferences a qualitative research method was used by looking at students' (B1 & B2 level) essays which have been marked and graded with the CEFR Linked Austrian Assessment Scale (CLAAS) by teachers who underwent rater training and have been working with it. The participants were teachers from the same school in Lower Austria who agreed to answer questions and provide copies of graded student writings. A questionnaire about personal experience and preferences to receive information about the raters' prioritisation was completed by the participants. The key research foci include investigating how the raters' prioritisation of certain textual features on the grading of extensive writing pieces is noticeable in their corrections, comments and written feedback in general. Furthermore, the study examines which categories the raters mostly prefer and how this could relate to their experiences with grading and being graded, their attitude towards the rating scale and their use of the scale. The categories in the scale include task achievement, organisation and layout, structural and lexical range, and structural and lexical accuracy.

The research question is whether teachers' personal experience, preferences, and beliefs about good writing still influence their decision-making process even when they are familiar with the analytic rating scale and underwent some kind of rater training. These influences become noticeable in the teachers' marking as they place more value on one dimension of the scale instead of considering them as equally important, for example by counting an error in more than one criterion. Analysing and interpreting the outcomes should help to discuss the implications for teaching and testing English as a foreign language, rater training, and objectivity in the process of grading. Problems might arise if there are no obvious preferences or the written feedback on the tests is limited or unclear, or the teachers use they scale correctly and grade the texts objectively.

The paper starts with an overview of the previous research in connection with rating scales and raters. Subsequently, writing in general and writing in the EFL context are discussed, teaching, testing and assessing writing are focused in order to establish a basis for the following topics of scoring methods, rating scales and the Austrian school-leaving exam, its English writing part and the rating tool used. Afterwards, rating related issues in connection with the raters and the rating scale itself are presented and discussed. Then the issues of bias, subjectivity and unfairness are explained and connected with rater errors as a source of bias. The impact of the rater errors on assessment principles completes the theoretical part of this paper. This is then followed by the introduction of the empirical part. First, the methodology is elaborated; second, the questionnaire, which was used to elicit information, and its basis are presented. Third, the participants and their backgrounds are introduced. Fourth, the results and their analysis are shown and discussed, and the implications for teaching and testing are presented. Finally, limitations of the study are revealed.

In order to examine the teachers' written feedback and the influence of raters' experience or bias on the rating process, the past work in this field of interest and recent developments must be considered. Introducing assessments scales in the context of language testing was an essential development and brought about a change in assessing the written work of language learners in Austria.

A number of research has concentrated on the differences between the various types of rating scales for the assessment of writing, mainly focusing on the advantages and disadvantages of using such scales (Brown & Abeywickrama 2010), their practicality (Hughes 2003), or the differences between the methods of assessment in terms of their formal and procedural aspects (Goulden 1992; Weigle 2002; Barkaoui 2010). Hughes (2003) provides teachers and raters with a guide to creating a test which fulfils the already established and proven criteria of a test system and the developing of a test. In his work, he brings together the main ideas and common principles that have been established in the EFL context over the past decades and, first of all, explains why these standards prove to be valid and useful. Secondly, he states how teachers can implement these principles in their own testing and test developments (Hughes 2003). In terms of testing writing, Hughes follows three main "rules": the tasks must be similar to the writing tasks which the students have to do in class on their language level, the tasks should actually give the pupils the chance to show their language abilities, and the scoring must be reliable and valid (2003: 83). Scoring written compositions with a rating scale is found useful and necessary by Hughes as the backwash is great when the scales are accessible to the students (2003: 105-6). Scales have to be adapted to the purpose and test specifications, but using a holistic or analytic scale for scoring writing increases a test's reliability, validity and hence its usefulness (Hughes 2003: 105-6).

In connection with the rating process using analytic scales, many researchers have used think-aloud protocols to gain insights into the decision-making procedure (e.g. Barkaoui 2011; Cumming, Kantor & Powers 2002). Barkaoui (2011) argues that although think-aloud protocols are useful for studies if the participants are not influenced or confused when they record, the marking of many teachers is altered and thus the protocols might not represent their actual rating process, especially the intuitive reactions, which are difficult to describe in words. Baker (2012) used write-aloud protocols to examine the differences between individual raters and their decision making in order to show that differences in cognition could explain rater variability. Other studies concentrated on the rating process itself, for example Lumley, who conducted a study to find out how a rating scale is used by teachers and in what ways their own marking process and general impression collides with the wording of the criteria in the scale with the result showing that "[...] the relationship between scale contents and text quality remains obscure" (2002: 246). His observation in relation to the process of assessing written compositions in a test is that although there is a valid and reliable rating scale for grading texts, the scale must be interpreted by raters and there often lies the problem: as raters have to apply the descriptors to numerous different texts, which is often a difficult matter, the raters do not know how to combine their impression of the text with the scale to come to a fair and objective score (Lumley 2002: 263). Lumley found out that raters use their feeling about the text in such situations more than the scale itself but still refer to its phrasing when explaining their marking (2002: 263). Additionally, Lumley is convinced that the raters in his study who had undergone rater training knew how to interpret the scale but still their methods of working with this tool seem dissimilar in many cases (2002: 263). Thus, he draws two revealing conclusions about the rating process. First, his research confirmsthat rater training mostly leads to no new way of scoring by the teachers but to learning how to use the phrasing of the scales to explain and justify their own impression of the writing (Lumley 2002: 267). Secondly, the centre of attention in this process are the raters, not the scale, as they still decide what is essential to them in the framework of the scale and give reasons for their choices (Lumley 2002: 267).

Cumming, Kantor and Powers (2002: 90) dealt with the raters' decision-making during the marking process in general and found out that any framework has its limitations due to "the complexity of human assessment processes, the interrelationships of language, ideas, and rhetorical forms, and the difficulty of distinguishing [...] discrete human behaviours from their holistically integrated nature".

All of these findings are extremely relevant and interesting in the context of the research at hand; a special focus is on the issue of the assessment of writing in the EFL context in general, its possibilities and limitations, the ways in which teachers function as raters and in what ways their decision-making process is influenced by their own experiences. These insights show more clearly which fields of interest are still underrepresented in the research of this field but might provide important insights into the topic and its relevance to the EFL classroom and assessment.

2. Writing

Writing is such an integral part of our lives, especially in times of texting, tweeting or blogging, that we rarely question its significance in everyday situations. However, in the context of writing in English as a foreign language and the interest of this paper, definitions of writing and its dimensions must be discussed to fully comprehend the extent of the debate on scoring and the implications of such discussions for the teaching of EFL writing.

In general, writing is a broad concept which can range from taking short notes to writing longer compositions (Weigle 2002: 7). In the context of the language classroom, intensive or controlled writing is defined by Brown and Abeywickrama as writing tasks that only ask for short and precise answers in order to show students' abilities to forming correct sentences or using a word correctly (2010: 267). In contrast, extensive or responsive writing includes "a continuum of possibilities ranging from lower-end tasks […] to more open-ended tasks such as writing short reports, essays, summaries, and responses, to texts of several pages or more" (Brown & Abeywickrama 2010: 275-76). The extensive writing tasks are of great interest in the context of this research as the *Zentralmatura* (the Austrian school-leaving exam) and other writing tasks on a B1 or B2 English level require the learners to compose texts in response to various prompts and write at least one longer text. Writing does not only fulfil the purpose of language learning in our context, it is also a "culturally recognised purpose, reflecting a particular kind of relationship, and acknowledging an engagement in a given community" (Hyland 2003: 27). Thus, language proficiency includes writing as a means of communication in the cultural context, which seems especially necessary if English is considered to be a world language, a lingua franca, the language of globalisation, the language of international business connections and the internet. Considering that globalisation is an ongoing process and that technological progress is evolving rapidly, the significance of the English language is increasing simultaneously with transnational developments, the growth of social media and technological advancement in the 21st century. This is also why writing is one of the major language skills one needs to possess in order to successfully communicate in English and make use of the language's status as a world language.

2.1. Writing in the EFL context

Having established that writing in English is a multifaceted, broad topic whose importance has been growing due to the increasing need to communicate on an international level, it is necessary to focus the process and the product of writing. Firstly, the subject matter of writing as a process is discussed and secondly, the product of writing in the context of language learning is dealt with in connection to rating written compositions and rating issues.

An essential difference must be highlighted: the difference between English as a second and English as a foreign language. Hyland explains that the former includes areas where English is spoken by the public and the latter includes areas where English is not one the official languages or not used by the public (2003: xvi). In Austria, English is taught as a foreign language as it is not one of the official or administrative languages.

As a start, a line has to be drawn between the writing of skilled writers in the L1 and in the L2. Hyland (2003: 36) lists the most striking differences between the two:

- General composing patterns seem to be largely similar in L1 and L2.
- Both L1 and L2 skilled writers compose differently from novices.
- Advanced L2 writers are handicapped more by a lack of composing competence than a lack of linguistic competence. The opposite is true for lower proficiency learners.
- L1 writing strategies may or may not be transferred to L2 contexts.
- L2 writers tend to plan less than L1 writers and produce shorter texts.
- L2 writers have more difficulty setting goals and generating material.
- L2 writers revise more but reflect less on their writing.
- L2 writers are less fluent, and produce less accurate and effective texts.
- L2 writers are less inhibited by teacher-editing and feedback.

Figure 1. L1 and L2 writing (Hyland 2003: 36).

These differences must be considered in the teaching of writing in the EFL classroom and it cannot be assumed that the L1 and L2 are learnt or improved in the same ways. Teachers need to make their students aware of these differences and provide them with new techniques to improve their writing skills. Generally, teachers should not assume that students know what good writing includes, as mentioned in the differences above.

As a communicative necessity in EFL learning, writing requires five types of knowledge according to Tribble (1996:11). These types of knowledge are necessary for the L2 writer to produce “effective texts” and must also be considered in teaching, namely content knowledge, system knowledge, process knowledge, genre knowledge and context knowledge (Tribble 1996:11). Tribble explains that content knowledge must be part of the learners’ writing repertoire to give their texts substance and to be able to express opinions on various subjects and issues (1996:11). As it is essential to know what to write about, reading, listening and classroom discussions are necessary to acquire content knowledge (Tribble 1996:11). Another crucial aspect is system knowledge, which refers to the language system as a whole, its grammatical rules, a wide range of vocabulary and its conventional codes or formalities (Tribble 1996:11). Process knowledge, which will be discussed in more detail below, includes acquiring strategies and methods to complete writing tasks successfully (Tribble 1996:11). Hyland (2003: 27) emphasises that genre knowledge helps the learners to realise the purpose of a text and how it differs from other text types, i.e. it contains the “communicative purpose of the genre and its value in particular contexts”. Finally, according to Hedge, context knowledge is concerned with writing “reader-based” texts which means that the writers must always be aware of their audience and how to create a text that is easy to read and understand and does not let the reader do the work of making sense of its content (2000: 307). The reason behind that, Hedge claims, is one of authentic language use as most texts which are written in real life serve a particular purpose and are aimed at a specific audience (2000:307). Weigle too argues that language learners must be able to adapt and vary their writing according to the social context in which their texts are produced (2002: 19). It is reasonable to assume that these five types of knowledge are an essential and useful base for understanding the writing process as a whole and can help teachers, as well as students, understand how writing can be improved. Therefore, gaining the five types of knowledge makes sense for every kind of writing task in the EFL context and can certainly be applied to the writing context of this paper’s study, i.e. writing on a B1 and B2 level in the Austrian EFL classroom.

Many scholars who are concerned with teaching writing in the EFL context agree that it is crucial to concentrate on the process of writing itself and what it means for the student (e.g. Hyland 2003; Hedge 2000). Hyland describes the process of writing as being essential for teaching: the language teacher must be aware of the process and the problems that young and inexperienced EFL writers face. The focus is not on writing being a linear approach but an ongoing process which is planned, revised, edited, changed and reworked continuously (2003: 10-11). As especially young writers often focus on the content and do not see the opportunity for continuous work in this field in order to improve their writing skills, the language instructors must consider the students' lack of knowledge of the numerous options to improve their writing during the process and familiarise them with the practice (Hyland 2003: 10-11). A widely used and acknowledged model for this writing issue is the original planning-writing-reviewing framework by Flower and Hayes, which is also used by Hyland to illustrate the importance of the process (Flower & Hayes 1981, Flower 1989 cited in Hyland 2003:11).

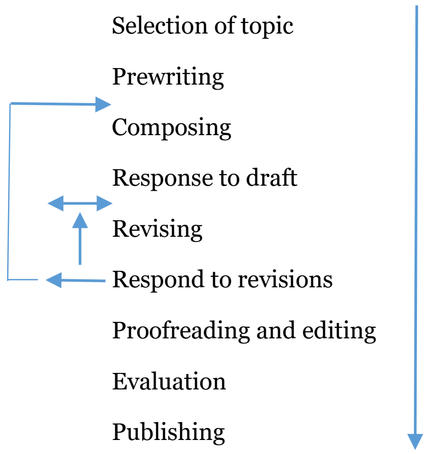


Figure 2. A process model of writing instruction (Hyland, 2003: 11).

Noting that this model allows writers to go back and forth in the process as much as they need to can help them produce a text effectively (Hyland 2003: 11). Hedge also puts emphasis on the teaching of writing as a process as there are many people, especially the language students themselves, who do not take into account that writing means planning, revising, editing and rewriting texts (2000: 302). In order to acquire knowledge of how to organise writing, she argues that certain teaching strategies with a focus on the process are necessary "to gain greater control over the cognitive strategies involved in composing" (Hedge 2000: 308). Similarly, Hyland (2003: 10) explains that teachers must focus their teaching on the techniques of acquiring and optimising writing strategies that include methods of organising to "develop students' abilities to plan, define a rhetorical problem, and propose and evaluate solutions". Therefore, the metacognitive level of writing has gained more importance and the teacher's role includes guiding the learners, providing them with various strategies and helping them find the methods and strategies that fit their learning styles (Hyland 2003: 11-12).

As the general importance of writing in the EFL context was shown, the significance of dealing with the procedural steps for producing good texts was discussed and the complexity of the process of writing was elaborated, the question of how the teachers' assessment can or should be adapted is a crucial one. This is a particularly interesting issue since writing is a part of EFL assessment which contains multiple facets and whose process is still not fully comprehended (Lumley 2002: 1). In general, it does make sense to treat texts in the classroom as a preparation for real-life communication and to maintain the process approach to writing. Nevertheless, there are situations in which learners have to write under pressure with the purpose of being assessed by their teachers. These situations demand that the students are well prepared and that their texts fit certain criteria and expectations, which might not necessarily correspond to the strategies and techniques they acquired for writing in a process-oriented writing classroom (Hedge 2000: 317-19). Therefore, exam preparation must be a part of writing instruction.

2.2. Teaching EFL writing

As already mentioned in section 3.1., writing includes various elements and stages on the way to reaching a proficient EFL level. It is a productive skill which students need to work on continuously, especially learners of English as a second or foreign language. The reasons for that are the many cognitive actions and ongoing developments during the process of learning a language, according to Hyland (2003: 11-12). In order to become a successful EFL writer, a learner must continuously work with the language and the elements of the writing process intensely (Hyland 2003: 11-12). Raimes (1983: 6) defines the elements of the writing process by portraying them in the following mind map overview:

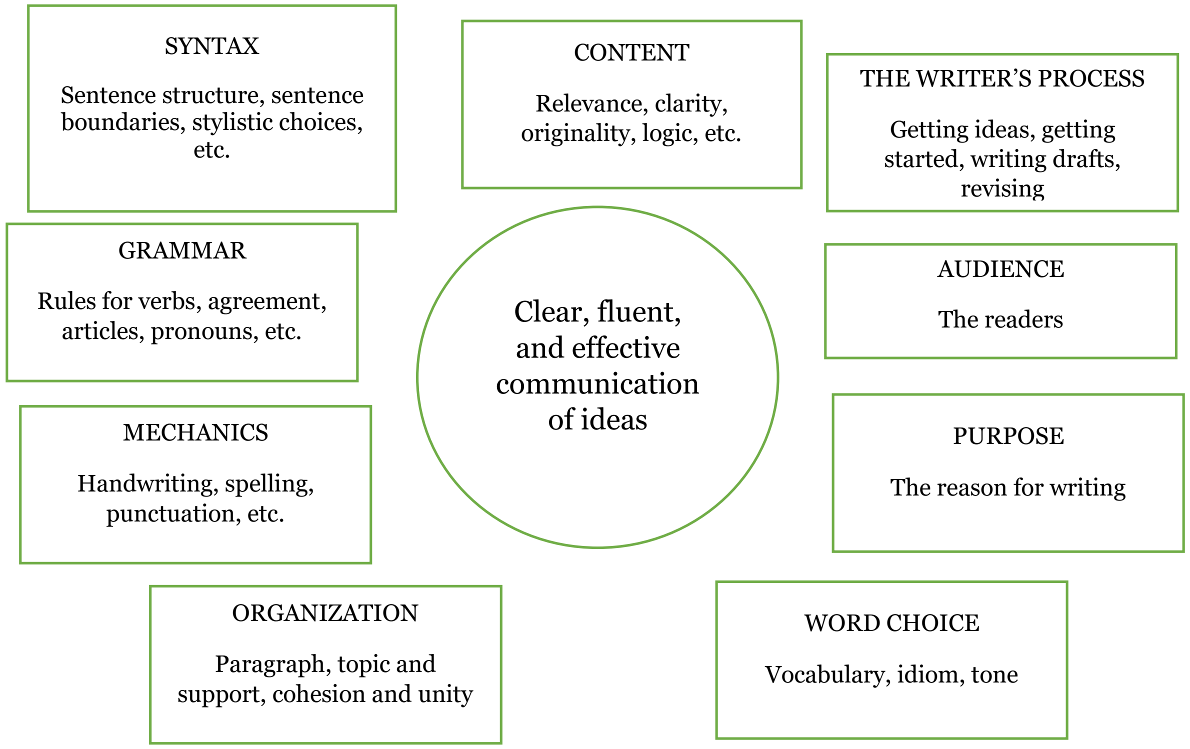

SYNTAX
Sentence structure, sentence boundaries, stylistic choices, etc.
CONTENT
Relevance, clarity, originality, logic, etc.
THE WRITER'S PROCESS
Getting ideas, getting started, writing drafts, revising
GRAMMAR
Rules for verbs, agreement, articles, pronouns, etc.
Clear, fluent, and effective communication of ideas
AUDIENCE
The readers
MECHANICS
Handwriting, spelling, punctuation, etc.
PURPOSE
The reason for writing
ORGANIZATION
Paragraph, topic and support, cohesion and unity
WORD CHOICE
Vocabulary, idiom, tone

Figure 3. Elements of the writing process (Raimes 1983: 6).

These elements again show the complexity and convoluted nature of writing and present the various facets of the process which have to be mastered. This illustration helps to show the difficult tasks both EFL learners and teachers are faced with. Language instruction and teaching writing successfully cannot be realised without constant work; assessment and feedback play essential roles in this matter.

The parts of writing are not only relevant to the writing process but they also play a crucial role in the assessment of writing. The elements in figure 2, except for the writer's process, are part of various rating scales as they also represent the essential features of writing. In the context of this paper and the interest in the CEFR Linked Austrian Assessment Scale (in short CLAAS), clear parallels can be drawn from Raimes' model to the criteria of the scale (BIFIE 2014a: 8-9). Grammar, syntax, mechanics and word choice can be found in the CLAAS criteria "lexical and structural range" and "lexical and structural accuracy". The features of organisation in this model can be found under "organisation and layout". Purpose, audience and content are similar to the criterion of "task achievement" (BIFIE 2014a: 8-9). The writing process and the assessment of writing, therefore, consist of similar components which are essential to both of them. Writing is not only difficult or problematic from the learners' point of view, but also from the raters' viewpoint, which is discussed after gaining insight into assessing writing in general and presenting an overview of the most prominent scoring methods.

3. Assessing writing

A discussion of the assessment of writing can only take place after listing the fundamental theoretical elements which are important in the complex area of testing and grading English. The circumstances for using an analytic rating tool in the Austrian EFL classroom are described and the rating scale is presented. Additionally, the research questions of this paper and a theoretical framework of teaching and testing principles can then be connected.

3.1. Principles of writing assessment

To classify the kind of test for which Austrian students are prepared and teachers use the rating scale in question, a closer look at assessment purposes and principles is necessary. Being able to measure students' abilities requires an understanding of which function different tests serve, how tests are designed and what kind of foundation tests are built on.

Test types follow different purposes for assessing the candidate's skills. Bachmann and Palmer (1996: 19) note a general difference in the function of teaching and testing: the foundation of teaching and classroom instruction is "to promote learning" whereas "the primary purpose of tests is to measure". That does not mean that tests cannot function as a pedagogical tool, but it is not their main aim (Bachmann & Palmer 1996: 19). How various test types differ in their purposes is illustrated by Brown and Abeywickrama (2010: 9) with the examples of "commercially designed and administered tests" and "classroom-based teacher-made tests". The former are most suitable for assessing levels of competence, which means that the students' "overall ability" is measured (proficiency tests) (Brown & Abeywickrama 2010: 11). The latter includes surveying the progress of the learners and checking whether the aims of the class have been reached or if any elements must be revised or discussed in more detail in the context of achievement tests (Brown & Abeywickrama 2010: 9).

When the school-leaving exam in Austria changed from an exam prepared by teachers, based on the contents and foci in their classes and the official curriculum, to a test that examines whether the candidates have reached B2 level, also based on the curriculum but independent from different focal points, by testing the same topics, text types, grammar, etc. for all students, the English *Matura* was changed from an achievement test to a proficiency test. The difference between the two lies in their purpose: an achievement test wants to find out whether a learner has acquired the contents of a class, course or similar and the exam is about a defined subject area covered in a particular time (a semester, a school year, etc.). The test specifications for an achievement test are, on the one hand, based on the work done in class, the aims of the class, the task types practiced during the lessons and appropriate timing (Brown & Abeywickrama 2010: 9). This is done to help students work on their weaknesses and tasks might vary due to the varying emphasis put on different areas (Brown & Abeywickrama 2010: 9). On the other hand, as teachers must follow the curriculum, the tests are based on the descriptors of this educational program (Brown & Abeywickrama 2010: 9). In contrast, a proficiency test measures the long-term language competence of the candidates and is not restricted in its content (Brown & Abeywickrama 2010: 9-11). According to Brown and Abeywickrama, tests are either compared to a norm or criteria (2010: 8); in the case of the writing in the *Zentralmatura* they are compared to criteria (i.e. B2/ CEFR), and provide summative feedback. Their score often includes only one grade and can be subdivided into categories that reflect the different parts of the test (Brown & Abeywickrama 2010: 9-11).

Knowing the purpose of the test helps teachers and students understand the test objectives and prepare accordingly. Moreover, there are main principles described by Bachmann and Palmer that can be applied to any test type; they are crucial for exams to be fair and comparable (1996: 19). Reliability, validity, practicality, authenticity, interactiveness and washback are the fundamental features of designing language tests. Two of these six principles are absolutely crucial for testing, namely reliability and validity; they are also called "essential measurement qualities" since they are the basis for rating through which the raters draw conclusions for applying scores (Bachmann & Palmer 1996: 19).

Test reliability is described as "consistency of measurement" by Bachmann and Palmer (1996: 20). Brown and Abeywickrama (2010: 27) explain this consistency by saying that "if you give the same test to the same student or matched students on two different occasions, the test should yield similar results". For a test to be reliable it must be dependable in the sense of containing explicit instructions and transparent scoring criteria which include a coherent set of descriptors. The exam and the system behind it must provide a framework for the rater to evaluate the test according to the scoring criteria, the test has clear test items and one way of testing its reliability is that it has to be “consistent in its conditions across two or more administrations” (Brown & Abeywickrama 2010: 27).

An exam is valid when it assesses “exactly what it proposes to measure”, when there is a theoretical foundation behind it, when it measures the candidate’s performance only and his or her competence can be shown through the rater’s selection of tasks which reflect the test’s objectives (Brown & Abeywickrama 2010: 30). Additionally, there should not be any unrelated items or unknown quantity in the test (Brown & Abeywickrama 2010: 30). A test is practical when it is relatively easy to conduct, meaning that there are no extraordinary measures taken to test what the teacher wants to test, and resources are available (Brown & Abeywickrama 2010: 27). Authenticity is concerned with a test’s relation to reality and actual language use, which means that the test content, i.e. its language and items, must be put into a context, follow some kind of logical system and can be used in real-life situations (Brown & Abeywickrama 2010: 37). Interactiveness is part of the main principles for Bachmann and Palmer (1996: 25) and they describe it “as the extent and type of involvement of the test taker’s individual characteristics in accomplishing a test task”. That involves the student’s language skills, their level of awareness of the subject matter, and “affective schemata”, which is explained by Bachmann and Palmer as “affective or emotional correlates of topic knowledge” (1996: 65). Interactiveness is often omitted in the discussion of assessment principles or addressed in the context of authenticity, e.g. Brown and Abeywickrama (2010: 36-37). However, Bachmann and Palmer emphasise that interactiveness must be treated separately from authenticity since a test’s authenticity depends on the agreement of the tasks in the test and the tasks in the target language use (1996: 25-26). Yet interactiveness depends on the agreement between the student who takes the test and the task itself. Therefore, the principle of interactiveness is characteristic of any task (Bachmann and Palmer 1996: 25-26).

The next principle, washback, is given when teachers, as well as students, can benefit from the test regarding the teaching and learning contents and methods (Brown & Abeywickrama 2010: 38). Moreover, the test takers should know how to study for the test and the circumstances should allow them to accomplish a high score. Positive washback can be achieved when the candidate’s language skills can be improved by the scorer’s assessment and comments, which should be more formative than summative (Brown & Abeywickrama 2010: 38). The difference between formative and summative feedback or assessment is that the former gives students information on their abilities during the process of learning how to compose a text, for example. In contrast, the latter does not give information about the same matter during the learning process but at the end of the class, course, etc. (Brown & Abeywickrama 2010: 7).

These principles are considered and applied by the test designers of the new standardised *Zentralmatura* in Austria and thus do not all have to be questioned in this context. Nevertheless, the principles of reliability and validity raise some issues for the Austrian *Zentralmatura* and the assessment of the writing part. It is important to note that these two cannot be discussed separately since "a test should do what it is intended to do and it should do it consistently" meaning that a test is only useful when both principles are observed (Hyland 2003: 215). Reliability does not only include the aforementioned conditions but can be subdivided into different aspects. Alongside student-related reliability, test reliability and test-administration reliability, there is so-called rater reliability (Brown & Abeywickrama 2010: 28). This includes intra- and inter-rater reliability, the first one being concerned with issues an individual rater might face during the scoring process like bias, having trouble interpreting the rating criteria or not paying a sufficient amount of attention. The second one means the extent to which teachers mark tests and come to a similar conclusion to other teachers, resulting in the same score (Brown & Abeywickrama 2010: 28). Both aspects of rater reliability can be improved by rater training and distinctive analytical rating criteria can enhance a test's reliability in this aspect (Brown & Abeywickrama 2010:28). Validity is another principle that is of importance in the discussion of the assessment of written compositions and related issues. Validity can be divided into several categories: content-related validity, criterion-related validity, construct-related validity, consequential validity (impact), and face validity (Brown & Abeywickrama 2010: 30). Content validity means that a test includes tasks which aim at measuring what they propose to measure. If the content validity of a test is questioned, the test specifications must be checked to identify which abilities or competencies they comprise (Hughes 2003: 26-27). Criterion-related validity in the EFL classroom can be proven by measuring the criterion not only in a test but also in another way to show that the results for the criterion correlate (Brown & Abeywickrama 2010: 32). In order for a test to have construct validity, the tasks must be based on previously elaborated theoretical frameworks which tell the test designer how to measure the skills that are to be measured (Brown

& Abeywickrama 2010: 33). Constructs are the conceptual mechanisms or operations that are included in what is then called good writing (Hyland 2003: 218). Consequential validity, also known as impact, refers to the outcome of testing in relation to the question of whether a test could assess the abilities it claims to assess, the consequences of the test for the candidates and also the purpose and perception of the test and the scores (Brown & Abeywickrama 2010: 34). Face validity describes a part of consequential validity in that it refers to the way the test takers perceive the test: whether it makes sense to them why they take the test or whether they consider it fair (Brown & Abeywickrama 2010: 35). For the purpose of this research, construct validity is essential as problems can arise in this area in the rating process, which is discussed in more detail below. A writing test can only have construct validity if the task "actually tap[s] into the theoretical construct as it has been defined" (Brown & Abeywickrama 2010: 33). Thus, in the case of writing in the EFL context in Austria, the construct of the *Zentralmatura* includes every feature that is regarded as being part of good writing (e.g. lexical range and accuracy, coherence, etc.).

3.2. Scoring methods: rating scales

In the assessment of writing, as in the assessment of other language skills, the most important aspect is to predetermine the objectives of the test and have a concept of what kind of text learners have to compose. As mentioned before, the task must be reliable, valid and practical. The range of methods for assessing a text varies considerably depending on the task type, the test type, and institutional regulations, and finding the most suitable assessment criteria can be a difficult task (Brown & Abeywickrama 2010: 259-60). To relate analytic scales to the study's aim of finding out whether this type of tool helps raters avoid subjective grading and bias, the alternative rating scales must also be discussed.

In general, rating scales help raters with their assessment of the features of a text in a process where "the language produced is compared to a scale descriptor for assessment" (Bukta 2014: 52). The scales' originality, their organisation, the fact that there are also zero points and their "equal intervals" are descriptive characteristics of this kind of assessment tool (Bukta 2014: 53).

For the scoring of writing tasks, Brown and Abeywickrama mention three methods: analytic, holistic and primary trait scoring (2010: 283). The first method for scoring written compositions that must be mentioned is analytic scoring as it is the most suitable way of giving detailed information on the test takers' skills according to Brown and Abeywickrama (2010: 284). This method provides washback to an extent that cannot be achieved with holistic or primary trait scoring as these methods are either too superficial or too narrow for the learner to gain insights into their strengths and weaknesses in every aspect of writing (Brown & Abeywickrama 2010: 285). Analytic rating scales treat the features of writing separately and require the rater to score each error in only one category which has to be predetermined. They "require the rater to provide separate ratings for the different components of language ability in the construct definition" (Bachman & Palmer 1996: 211). According to Bukta, this type of rating scale is usually made of several categories which are based on theory-bound criteria and relate to the language learners' proficiency stages, and the scales are divided into categories such as the ones in the CEFR Linked Austrian Assessment Scale (CLAAS) discussed below (2014: 75). This approach is beneficial for improving students' proficiency as the individual scores on the different writing criteria or dimensions display more clearly which issues students might have in certain areas and in which of the categories they prove to be competent or skilled (Brown & Abeywickrama 2010: 285). A detailed set of scores in a writing task is practical for students to make sense of what to work on and might also give them the motivation to work on their weaknesses when they can see that they are strong in other categories. However, at the same time, analytic scoring is more impractical in comparison to the other types of scoring (Brown & Abeywickrama 2010: 285). The scoring process takes longer as the raters have to apply the descriptors of the detailed scoring scale to the texts, which means that usually, the texts also have to be read several times (Brown & Abeywickrama 2010: 285). Weigle also lists several advantages and disadvantages of analytic rating scales and concludes that, on the one hand, they are reliable, students gain more information on the development of their writing skills and detailed feedback can be given on their strengths and weaknesses (2002: 121). On the other hand, she agrees with Brown and Abeywickrama and claims that analytic scales are not the most practical assessment tools as it takes the rater more time to work through all of the many descriptors in the different categories and reading a text the way it is necessary to do while rating analytically does not seem to be authentic (Weigle 2002: 121). Nevertheless, Brown and Abeywickrama argue that this rating approach is beneficial for the test takers as it gives them more insight into the areas of writing which they need to work on but also shows them the areas which they succeed in (2010: 284).

Holistic scoring requires raters to get an overall idea of the writing and rate the text by comparing this general impression to predetermined descriptors. This assessment technique is not only holistic in the sense of looking at a general impression of writing skills but also regarding the score, which is not composed of a set of points in different categories (Brown & Abeywickrama 2010: 283). Although, as Weigle states, holistic scoring is rather a quick method of assessing written work, inter-rater reliability is likely and reading a text holistically seems more authentic, there are some noticeable drawbacks (2002: 114). First of all, Weigle (2002: 114) argues that "achieving high inter-rater reliability at the expense of validity" is a problem in connection to holistic scoring. Secondly, as there is only one final score, the learners do not receive a mark on the elements of writing the score naturally contains, which might lead to a lower score for students whose writing performance shows strengths in one area which are overshadowed by their weaknesses in other areas (Weigle 2002: 114). Similarly, Bukta (2014: 76) compared the holistic approach with the analytic one and found out that "a holistic scale is built on an assumption that writing ability develops evenly, whereas an analytic scale can make a distinction between the elements of ability". This indicates that holistic scoring does not give detailed feedback on the components of writing which could cause a lack of information on how to improve their writing for the learners. Weigle (2002: 121) agrees and argues that "different aspects of writing ability develop at different rates" which emphasises the advantage of analytic rating scales as a feedback and washback tool. Thirdly, raters must be trained intensively for several reasons. Vaughan states various ways of assessing texts holistically which might complicate the usefulness of such a scale (1993: 118;120). There is, for example, the "first impression dominates approach" or the "grammar-oriented rater" (Vaughan 1993: 118;120). In cases where raters are not entirely sure how to relate the descriptors of the scale to the writing piece, the risk of trusting their own judgment is relatively high or the raters' own notion of what is most important determines the outcome (Vaughan 1993: 118;120). Elbow also argues against holistic scoring for classroom use due to this scoring method being too open to interpretation and beliefs about what a good text must look like from the teacher's point of view (1996: 121-22). Elbow (1996: 121-22) also warns against trusting statements about reliability as he argues that the "profession's solution" has been that many people only adjusted their scoring methods to each other and trained the raters to use the same marking foci for texts to be graded similarly. Such an approach does not make much sense since there would be many different teacher groups who all concentrate on different aspects of good writing, depending on what kind of training they underwent and what their focus of rating is. Thus, there would not be a common understanding of what good writing looks like or how students can improve their skills to create successful writing compositions. This is particularly difficult in a setting where standardised tests are used or introduced, as reliability and validity cannot be guaranteed by conducting standardised exams alone but the tests also have to be corrected and graded with a rating system that can be applied and understood by the teachers involved. Ideally, the rating scale would not leave much room for interpretation and can easily be applied by trained raters.

Another assessment method for L2 writing is primary trait scoring, which concentrates on the task and its aims and assesses whether the candidate manages to reach the most salient goals of the prompt (Brown & Abeywickrama 2010: 284). That is, if the task was, for example, to persuade someone to do something, the writer would be assessed only for being persuasive. Organisation, lexical and structural range, the accuracy of vocabulary and grammar and other textual elements that amount to a good text are inevitably part of rating with the primary trait system as these features must be given for a text to be successful. When practising the function of a text, this approach is beneficial as it gives the teacher and students an insight into the specific methods of fulfilling a task depending on its purpose (Brown & Abeywickrama 2010: 284). If students need feedback on the underlying elements of their writing, such as structure, lexical range or grammatical accuracy, or information on how their writing is progressing, this approach is not helpful due to its limitation to only one specific aspect of writing, depending on the focus of the task (Hyland 2003: 230).

In the context of the Austrian upper secondary and the school-leaving exam, an analytic rating system is used for the written English test. An elaboration of this topic follows in section 3.3.

3.3. The Austrian *Zentralmatura*

In this section, the Austrian *Zentralmatura* is briefly discussed to establish the context in which the analytic rating scale in question was created for and is currently in use. The new Austrian *Zentralmatura* was implemented in 2014 and fundamentally changed the school-leaving exam (BMBWF n.d.). From that time on, teachers were no longer the ones who selected or created the tasks for the exam; the

BIFIE (*Bundesinstitut für Bildungsforschung, Innovation und Entwicklung des Bildungswesens*) was commissioned to do that. This institution was established by the Austrian government in 2008 to manage educational standards, international assessments, national education reports and the development of the standardised school-leaving exam (BIFIE 2018).

Before the implementation of the *Zentralmatura,* the school-leaving exam had been criticised due to questions of the tests' reliability and validity and the lack of transparency since the teachers had been allowed to use any tasks in the test and a national comparison of the students' performance would not have been valuable. Similarly, the fact that the test in Austria is now standardised has often been criticised in the media in terms of practicality as well and the creation of tasks by the institution in charge. Among teachers, the implementation of the *Zentralmatura* had a difficult start but was widely accepted after some time, as a recent study has shown, and the more experience the teachers have gained with the new testing methods, the more positive their attitudes have become (Lopatina 2016: 60). Generally, the change has been perceived as a far-reaching one, causing a wave of discussion among the Austrian people and the media.

One aspect that has also been criticised since the implementation of the *Zentralmatura* is the fact that the grading of the more open tasks of the tests, like writing, is still done by the teachers of the candidates which, as will be discussed below, might influence the raters in their decision making or grading process. Nevertheless, the new Matura has brought more transparency, comparability, and fairness to the Austrian school system as not only the final exams are standardised but also the rating system has undergone some major changes. Accordingly, the tool, i.e. the CLAAS, was standardised which is used to correct and grade the written compositions of the students. Additionally, teachers are required to use these scales throughout upper secondary for grading written exams to familiarise the students with the demands and rating criteria they will face in their school leaving examination (BIFIE 2014a: 1-2).

3.4. The CEFR Linked Austrian Assessment Scale

The analytic rating scale which was developed for the scoring of the written part in the standardised English Matura in Austria is mainly based on the *Common European Framework of Reference for languages,* CEFR in brief (Council of Europe

2001). Other holistic or analytic scales were used in the past, but they did not reflect the language skills of a B1 or B2 level writer in English as a second language since these scales were developed before the CEFR and thus had to be replaced by the CLAAS to include many of the crucial descriptors the CEFR specifies. The elaboration of the new scale includes a description of the main principles and rules and recommends using the tool to correct homework and tests from a B1 level onwards, i.e. in the upper secondary classes in order for the pupils to adapt to the system and know what is expected of them in the final school leaving exam (BIFIE 2014a: 1-2; BIFIE 2014b: 1-2). There are two scales, one for the B1 level (BIFIE 2014b), which was published in German, and one for the B2 level (BIFIE 2014a), published in English. As the two versions are similar in their content and only vary in their language level requirements but not in their explanations of guidelines and rules, the English version will be used and cited throughout this paper.

First of all, the scale includes four criteria which reflect the main features that a good text must contain: task fulfilment, organisation and layout, lexical and structural range, and lexical and structural accuracy. Every criterion contains eleven bands which include descriptors for every even band (BIFIE 2014a: 8-9).

The first criterion is task achievement (TA) which must indeed always be the first dimension to be marked as it includes a veto descriptor which means that if the candidate did not manage to write about the topic in the given task, the other three criteria must also be assessed with zero points. Besides that, this criterion deals with the length of the text, the topic, the content development and expression of opinions (BIFIE 2014a: 3). The second criterion is organisation and layout (OL) and encompasses all structural features at text, paragraph and sentence level. Organising ideas, considering the audience of the text and “the test takers’ awareness of different layout conventions for different writing tasks” are part of this dimension (BIFIE 2014a: 3). The third criterion, lexical and structural range (LSR), includes register and variation in lexis and grammar. In the fourth and last criterion, lexical and structural accuracy (LSA), grammatical and lexical correctness are rated and whether the text adheres to punctuation and spelling rules (BIFIE 2014a: 3). The sequence in which the criteria were described should also be the order in which teachers mark a text. Furthermore, it is suggested that the beginning of the rating process is always band 6 as this represents the pass mark or a minimally competent candidate. If all the descriptors apply to the performance in the category, the teacher moves on to the next band, 8, to see if these descriptors apply as well, and so on. The uneven bands in between, which do not contain descriptors, can be chosen if the rater is indecisive about whether to allocate, for example, band 6 or 8. The rater can then decide to allocate band 7 in the criterion if the descriptors from band 6 partly apply to the text but also some of band 8 (BIFIE 2014a: 4). The descriptors in the rating tool must be applied four times, meaning that every category must be applied "independently from one another" for which the texts must be read at least four times (BIFIE 2014a: 4).

The foundation for this assessment scale mirrors the previously discussed features of EFL writing by Raimes (1983: 6) as every dimension of the tool deals with different features of writing and tries to include all of them in a descriptive apparatus to support constant work on writing skills.

4. Rating-related issues: raters & rating scales

To analyse the process of rating and the issues that arise in the process, it must be clarified what and who a rater is, or which skills and knowledge a person has to have in order to become a qualified rater. In this section, general difficulties of the rating process are presented, raters are described, various rater related problems are shown and analysed, the options one has in order to become a qualified rater in Austria are (re-) viewed, and the connection between rater and rating scale will be presented. Then the wording of rating scales is briefly discussed and the issues of bias, subjectivity and unfairness are related to rating scales.

4.1. Difficulties in the rating process

Generally, in the process of rating students' writing, there are many challenges a rater has to respond to. There are some specific rater issues which will be discussed in section 4.3. or problems with the rating scale which will also be discussed in section 4.4. However, there are some general topics to be discussed first.

Teachers who rate texts according to descriptors and standards, in the case of Austria provided by *BIFIE's* assessment scales and guidelines, are influenced by either external or internal factors. For example, many Austrian English teachers have not undergone any rater training. Moreover, Cohen (1994: 308) remarks that the process in which teachers decide how to grade a text is always also characterised by their "assumptions, expectations, preferred rhetorical models, world knowledge, biases, and notions of correctness". Therefore, Hamp-Lyons suggested that these influences must be taken into account in the interpretation of the raters' grading and the underlying process (1991: 263). Considering these obstacles in the marking process, Bukta emphasises the importance of rater training to block out teachers' subjective perceptions of a text (2014: 113). Standardising tests and marking procedures makes it less difficult to train teachers to concentrate on a student's proficiency instead of on other factors that might influence them. Still, Bukta notes that constant decision making and the process of grading itself, from looking at a text for the first time to giving the final mark, must be paid extra attention; the descriptors, the test's objectives, and its construct as well as useful ways of reaching a final score must play an essential part in training teachers how to use scales for a standardised test (2014: 113).

Difficulties in the rating process might arise due to the raters' interpretation of the descriptors, either because of the wording of the scale or rater related issues. Moreover, some writing pieces might not be easy to relate to the descriptors or they do not match them at all. Additionally, the learners' or raters' views and beliefs towards the learning process can vary or might not correspond with the views presented in the rating scale (Upshur & Turner 1995: 5-6). Commonly occurring issues in connection with the raters will be discussed in the next section.

Although opinion is divided on the methods of assessing writing due to its complexity, teachers' preferences and different experiences, issues of bias and raters' own expectations of texts, researchers have been trying to establish rating systems that try to take all of the above into account and to find appropriate rating processes for various testing situations and purposes. In section 4.3. rater and rating issues will be discussed in order to connect these ideas and illustrate the complexity of this topic in more detail.

4.2. Raters: definition

In the context of assessing English for achievement tests, a rater is usually the teacher of the student who writes a text. In some countries in which standardised proficiency tests are common, the raters can also be other teachers from the same school, teachers from other schools in the country, or even people who work for the institutions where the tests are developed.

In Austria, the raters had always been the teachers of the students, regardless of the test situation, with only a few exceptions. The usual classroom exams during the year are graded by the students' English teacher; it would not be practical to have someone else correct them, except in cases where problems arise or a teacher is unsure what to do about the grading. In such cases, many teachers rely on the help of their colleagues or the guidelines in the assessment scales. The writing part of the *Zentralmatura* consists of two extensive writing tasks, with one being shorter than the other. Although the *Matura* was officially changed from a non-standardised to a standardised exam in 2014, the raters are still the candidates' English teachers, but the test can be looked at by a second rater, which is especially necessary when the result of a candidate's test is somewhere between *Genügend* or *Nicht genügend* (which means that the grade is between the last positive mark to pass and a negative grade). The guidelines set by the Ministry of Education determine that the students' work must be corrected with the assessment tools provided by *BIFIE* and then the class teacher has to hand the written exams over to the assigned chairperson who has to check and confirm the grades (BMBWF 2018). Afterwards, the grades must be decided by a board of teachers, the headmaster or headmistress and the chairperson.

Thus, raters in Austria are teachers of the students rather than trained raters. The rater training teachers receive in Austria in connection with any school subject is limited. Now most teachers have their first contact with the rating system for their subjects in didactic courses at university. Some departments, for example the English Department at the University of Vienna, offer introductions to the application of the rating scales, but generally, the topic is only addressed minimally and efficient rater training does not take place continuously. Older teachers might not have had any training at university as the use of the analytic rating scale was only introduced a recently. Apart from learning how to use the scale, teachers came into contact with various rating methods in their own time as a student when they were rated and graded by their teachers. This individual experience that influences the testing methods of teachers at university level, as Berger explains, can be related to teachers in upper secondary as well (2014: 2-3). English teachers in Austria have three more options to learn how to use the scales: either they attend advanced training provided by the *Pädagogische Hochschule* or university or they start working with the scales and try to apply the instructions and rules that are usually attached to the scales by *BIFIE*, or learn from their more experienced colleagues by rating some work together.

Until now, there are no other options for teachers to gain more experience and expertise in this area.

When training teachers how to use the assessment scales, two major problems arise: the wording of the scale and the grading process itself in which the rater faces several problems. The former issue will be discussed in section 4.4., the latter is quite obvious: since grading students' pieces of writing is never the same and results vary due to the countless ways of interpreting a task, the teachers' grading can vary unless, as Weigle points out, they undergo continuous training on how to interpret and apply the assessment tools (1998: 280-81). Consistency and intra-rater reliability are improved by the training (Weigle 1998: 263). In the next section, further situations will be discussed which complicate the rating process concerning raters and their grading practice.

4.3. Rater issues

Correcting written assignments and giving written feedback on the correction and grading process are tasks that every teacher has to do on a regular basis. There are numerous factors which influence the process, the rater being one of the (McNamara 1996: 3). Saal, Downey and Lahey (1980: 415-19) name five main complications that occur in the rating process: severity or leniency (1980: 417), the halo effect (1980: 415), central tendency, restriction of range (1980: 417), and a lack of inter-rater reliability or agreement (1980: 419), which will be discussed in the next section. These errors are then connected to bias, objectivity, and fairness in the grading procedure in general.

4.3.1. Severity or leniency

Student performances in written assignments are often not graded accurately and frequently leave candidates with lower or higher points than they should have received (Saal, Downey & Lahey 1980: 417). The main issue here is the tendency for raters to be either too severe or too lenient. Grading texts this way can have profound consequences: Congdon and McQueen (2000: 164) found that although raters might be consistent in their grading and maintain their degree of severity or leniency, if the rater is on either one end of the severity-leniency spectrum, this influence "can turn an assessment into a lottery". Nevertheless, it can also be argued that consistent grading, although it might be too lenient or too strict, is still a more preferable approach than grading without any consistency.

Engelhard (1994: 98) explains that severity can be “viewed as a continuum” and measured by having raters undergo “a calibration study in which a common set of student compositions are rated by multiple raters”. To compare the results, an expert group can be formed and used to establish how the written work should be corrected and which feedback is more lenient or strict by putting the most crucial aspects of raters’ severity or leniency on a scale (Engelhard 1994: 98). In his study, Engelhard (1994) found this rater error to influence the results and therefore it counts as an influential factor in the grading process which must be considered in the reflection and improvement of rater performance. Nevertheless, being aware of raters’ tendencies to be too strict or lenient helps to solve the problem as their ratings can then be calibrated and adjusted accordingly (Engelhard 1994: 108).

4.3.2. The halo effect

The halo effect refers to situations in which raters cannot differentiate between “conceptually distinct and independent aspects of a student's composition” (Engelhard 1994: 98). This includes cases where the rater must use an analytic rating tool, like the CLAAS scale in this study, in which four dimensions are distinguished and must be treated separately by not counting one mistake several times but scoring each mistake in only one of the categories without including one’s general impression of the text in the final grade; the same applies for allocating bands for the parts which are successfully written: positive aspects of the text can also only be counted once. When the rater fails to do that and rather applies a holistic rating method, the halo effect takes place. This might be visible by the rater constantly assigning the same band in every category of the scale (Engelhard 1994: 99). In the process, the raters identify a variable which they believe to connect all these different analytic categories and therefore influence the scores by lowering or raising them (Feeley 2002: 579). The halo rating error is very common and a shared belief among researchers is that it occurs whenever any rating of human work is done by people (Feeley 2002: 578).

This effect can be divided into the “true halo” and the “illusory halo”, which are terms established by Murphy and Cleveland (1991). The true halo effect occurs, for example, when a rater grades a text by assigning the ratee band four in all four categories of the CLAAS scale, but it turns out to be by accident as the same teacher does not usually grade essays that way but the ratee indeed scored four points throughout the scale. The illusory halo effect occurs when the rater awards a student the same band in all the dimensions in order for the composition to fit the general impression the rater obtained from the text. Regarding the halo effect with this distinction, the true halo is not a rating error and the scores seem to be a coincidence instead of a rater not using the analytic rating scale properly (Engelhard 1994: 99). For cases in which the good impression influences the rater's decision, the halo effect applies. For the opposite phenomenon, the so-called horns effect applies as it "is the concept by which a person who is judged negatively on one aspect is automatically judged negatively on several other aspects without much evidence" (Belludi 2010).

When a halo effect occurs, its causes must be detected for reflection and avoiding this error in future assessments. In his comment on the halo effect, Feeley (2002: 578) detects one main possible consequence of the illusory halo effect: as the raters does not distinguish correctly between the categories in the assessment scale, they see more connections between the rating dimensions than there actually are. This leads to a false image of the textual elements being connected by errors or the general assessing of the rater and wrongly depicts the scale's categories. Another negative outcome of the effect is the influence on validity as the error "lower[s] the discriminant validity of the obtained ratings while also reducing construct validity" (Feeley 2002: 579). In a wider sense, considering the number of halo errors and the assumption that this is an ongoing problem in the rating of data in various areas of education, important decisions, selection processes and occupational advancements could all be based on errors connected to the halo effect (Feeley 2002 578-79).

The causes of the halo effect stem from cognitive processes in the raters' minds and were researched by Lance et al. (1994: 332-33). Three models of the error were established: the first one is called the "general impression model", which says that an impression influences the grading of the rater but ignores the actual scores and their connection to each other. This general impression stays with the rater throughout the whole rating process and obstructs the analytical thinking and separation of categories that are necessary for a grading procedure (Feeley 2002: 579). The second model is called the "salient dimension model" and is described as a problematic connection made by raters when they transfer their opinion about a person's attribute to another attribute of the same person. Feeley (2002: 579) mentions the example of the good-looking outer appearance of a student which “halos the evaluation of other related or unrelated variables (e.g., intelligence)”. The third model, the “inadequate discrimination model”, includes the evaluator’s inability to isolate the candidates’ performance in one area from their performance in another (Feeley 2002: 579-80). In the context of rating texts with the Austrian analytic rating scales, that would mean a teacher is influenced by a student’s good or bad performance, for example in the lexical and structural range criterion, and therefore bases his or her scoring for lexical and structural accuracy on this previous perception of the student’s accomplishment.

4.3.3. Central tendency

The third issue from the rater’s perspective is “central tendency”, a term referring to the phenomenon of raters avoiding the low or high end of each category in order to stay in the middle of the overall points which they apply too often. The problem here seems to be an inaccurate use of the rating scale which might stem from a lack of rater training and the lack of knowledge about how to interpret and use the scale (Engelhard 1994: 99).

4.3.4. Restriction of range

The term “restriction of range” is often confused with the the issue of being too lenient or strict or the problem of central tendency described above. The issues are indeed related since the restriction of range refers to the general problem of raters concentrating their scores on one point in the continuum of the rating scale’s categories. This might be caused by the person who grades the writing having a tendency to rate a text more leniently or strictly, or, as discussed above, to avoid extremes and find themselves stuck in central tendency. In the cases of being overly strict or lenient, the issue is a restriction of range; in the case of a focus on the middle, it is called central tendency (Saal, Downey & Lahey 1980: 418). A consequence of this restriction issue is again the question of the existence or maintenance of validity in the ratings (Engelhard 1994: 100).

4.4. The wording of rating scales

One major issue that still needs to be discussed is also concerned with errors made in the rating process and with the application of assessment scales, namely the misinterpretation or difficulty of interpreting the phrasing in the scales. Berger (2014: 63) distinguishes two main problem areas in this context: “the *application* of the scale in practice or the *nature* of the scale itself”.

Firstly, applying the descriptors of the rating scale can be problematic as it might not be possible or easy to “interpret and apply the scale descriptors consistently because the rating scale may not allow a common interpretation” (Berger 2014: 63). This problem can partly be solved by training raters and thus increasing inter-rater reliability (Berger 2014: 63). However, Berger mentions that despite rater training, this issue remains problematic (2014: 63-64). A reason for that could stem from the raters’ own opinions about and preferences for certain textual or language features in the assessment tool and therefore their different focal points can result in different scores. That is one of the reasons why the questionnaire in this paper’s study asks for the participants’ opinions on various textual features to investigate their personal preferences which might influence their written feedback on their students’ test texts.

The scale itself can also cause issues in connection with circumstances which could undermine the assessment principles of being a reliable and valid instrument for grading. As Berger points out, the character of rating scales can be problematic since the development of rating scales could be based on the personal beliefs or assumptions of the creators of the scales and not on a carefully researched theoretical background (2014: 64). That does not mean that the scales might not be useful but “[t]he validity of such scales is typically proclaimed by the authority of the scale developers or users” (Berger 2014: 64) which might lead to the scales being invalid or unreliable. Furthermore, Berger notes that although analytic rating scales suggest that students’ language competence grows, a theoretical background must again be the foundation of this claim since assumptions cannot be made about how the learning proceeds (2014: 64-65). However, research on this topic provides detailed information about second language learning (Berger 2014: 64-65).

Considering these two points of criticism, it becomes more apparent why many raters face various problems in the rating process. Under closer examination the CLAAS is not an exception to these issues. Although the descriptors in the scale are based on the CEFR, which have been carefully researched, developed and, as Berger (2014: 66) describes them, “empirically validated”, their wording and the order of statements which illustrate the language learning and process can also be questioned.

In this context, the cause of several errors in the process can be, for example a rater's belief in correlations between separate categories due to the interpretation of certain terms. For example, a rater could interpret the descriptors of one category as being connected to the descriptors of another category, which can then lead to a result that does not distinguish the two categories and bands were assigned due to this overlap or one error led to the choice of a lower band in two categories. Additionally, Berger's research states that although the descriptors of a CEFR based scale might not be the most vital factor to question, the scale's relation to the language learning progress and practice can be challenged (2014: 66). The bands in the scale suggest writing skills and their progression from zero to ten with certain competencies students acquire as they improve their skills. However, whether this process illustrates actual language use and the process of improvement can be questioned (Berger 2014: 66).

Furthermore, a problem with the CLAAS scale is that the descriptions in the four criteria do not always strictly separate the categories and thus it can be hard for raters to distinguish the various textual features. Therefore, the wording of the scale and also the raters' interpretation play an essential role in the grading process and must not be underestimated as these factors contribute to the teachers' performance as a rater and their decision-making process.

4.5. Bias, objectivity, fairness

In this section, the previous elaboration of rater errors is to be connected to the overall problems of bias, objectivity and fairness. However, before this can be done, these issues must be defined and treated in more detail to establish the context and aim of the study that follows in section 5.

A very basic understanding of bias in the field of assessment is related to the principles of usefulness, namely reliability, more concretely rater reliability. The terms were discussed in section 4.1. and reliability is just as essential in the course of rating someone's written work as validity, which will be discussed below. Bias, as a form of unreliability, occurs when a rater fails to be impartial and favours someone or impedes somebody's progress based on, for instance, their status as a good or bad student (Brown & Abeywickrama 2010: 28). Bias might emanate from rater errors like the halo or horns effect, or the appearance of the person whose work is rated or their "assigned status such as being classified as gifted or learning disabled" (Malouff

& Thorsteinsson 2016: 245). Bias can also occur in relation to the task itself, as He et al. state (2013: 479). They examined rater bias in the EFL context of college students' essays and found out that even competent, experienced raters can be subject to bias in that they graded argumentative essays more severely than descriptive ones. Additionally, the study shows that the raters were more lenient with their scoring in the textual criteria than in the language use criteria, which indicates that they might be biased towards particular writing elements (He et al. 2013: 479).

Objectivity then refers to the state of grading a text without any bias or subjective impressions or feelings; whether this is even possible must be discussed in another context and cannot be answered clearly due to the complexity of the topic. To establish a degree of objectivity, Malouff and Thorsteinsson argue that with assignments like essay writing, where judging takes place as the answers can vary due to the openness and room for interpretation of the prompt or task, rater objectivity could be established by anonymising the assignments (2016: 246). For the scoring of tests which consist of multiple choice or true or false questions, where there is a clear distinction between right or wrong, namelessness is not necessary to create a fair scoring process since there is no room for interpretation and judgment is not necessary as long as the answer key is correct (Malouff & Thorsteinsson 2016: 246).

Both being unbiased and staying objective in the grading process are part of fairness in the assessment of writing. The three terms cannot be separated from each other and become interrelated. Brown and Abeywickrama mention fairness in connection with test bias and fairness and define it as looking at a test and being free from assumptions or prejudices, ranging from age to ethnicity, among other examples (2010: 96-97). Furthermore, a test is considered unfair when a group of candidates is given an advantage based on their cognitive abilities or background knowledge, except for tests whose construct clearly aims at testing candidates' cognitive abilities. Giving students such an unfair advantage might happen in all kinds of tests where particular task types are used which might cause a group of candidates more problems than others (Brown & Abeywickrama 2010: 96-97). An example of that would be the new *Zentralmatura* in mathematics in Austria. In this test, not only operational, logical and problem-solving strategies are asked of the pupils, but they are also asked to answer open questions by writing in full, meaningful sentences which explain a mathematical problem or concept. This has been criticised frequently in the past years, especially in the most recent test

(Taschwer & Illetschko 2018). Although this approach might be preferable since Brown and Abeywickrama (2010: 97) argue that the "multiple intelligences present within every student" must be encouraged, the question remains as to whether the way of testing these abilities is favouring or neglecting a group or if it is a fair way of testing for every student. If indeed their German writing is assessed is a maths test and it is not part of the test construct, the criticism of unfairness seems to be justified.

4.6. Rater errors as a source of bias, subjectivity and unfairness

Drawing on this knowledge of rater errors, there are many situations in which unfair rating occurs. As mentioned before, when a rater is constantly too strict or too lenient, fair grading cannot be guaranteed; instead, the opposite happens, and students' work is measured along an unrealistic and unknown scale. Similarly, when a rater grades randomly without any consistency, fairness is not guaranteed. The real standards of the assessment scale and their foundation are lost due to the falsely used parameters of the rating tool.

Another connection between rater errors, bias and objectivity can be seen in the halo effect: in the general impression model, the example of sympathy is mentioned by which teachers can be influenced in the grading process, preventing them from staying objective as raters. This might lead to a positive impression of the ratee on the rater being the crucial factor in scoring his or her written composition (Feeley 2002: 579). Thus, a fair rating of a written piece of work cannot be guaranteed when a personal feature or an emotional bond affects the evaluation of the ratee's writing. Not only is this grading approach unfair, it is not objective or unbiased either. Moreover, the consequence of the halo effect influencing important decision-making processes like giving scholarships, acceptance processes at universities etc. shows that the unfairness of this effect becomes even more far-reaching. For instance, candidates who were graded in the correct way throughout their studies, etc. might not get the job opportunities they want because of other candidates' advantage that comes from favouritism in their education, i.e. grades. Hence, the halo effect raises the crucial question of what the consequences are when a teacher fails to rate a text analytically and whether this person can give sufficient feedback about a learner's strengths and weaknesses for the learning development. Considering such a fundamental deficiency, a definite yes or no cannot be the answer to these questions. Furthermore, the halo effect in the sense of raters basing their decisions of assigning points for one category on the scores of another category in which the ratee did either well or badly is an inherently biased rating approach. The same conclusions apply for the above mentioned reversed halo effect, the horns effect, as this rater error influences the rating process in the same way as the halo effect with the only difference being that a negative impression affects the grading (Belludi 2010).

The problematic situations arising from central tendency cannot be viewed as a biased grading approach but are definitely unfair for the candidate whose text is rated in that fashion. When raters do not know how to apply a rating scale and do not achieve their goal of providing feedback for the student and in a wider sense also positive washback, then the fairness of the process itself must be questioned and therefore also the student's potential for future development. If the teacher cannot understand the scale and its use for providing a framework for different aspects of writing, the students will not be able to make this distinction and work on separate features in the writing process either. The same can be said about the issue of restriction of range because a teacher who refrains from using the range of assessment criteria for the evaluation of texts cannot be viewed as a fair rater and does consequently also not provide authentic and sophisticated feedback for his or her learners.

Clearly, these rater errors have a great impact on the rating process and test results as they lead to bias, and subjective, unfair approaches. In the next section, the influences of such errors and bias on the principles of assessment are presented.

4.7. Impact on assessment principles

In all of the above-mentioned errors and their consequences, three major principles of the usefulness of the assessment are violated: reliability and validity on the one hand and washback on the other. As mentioned in section 3.1., the two elements of reliability and validity are interrelated and must be given for a test and test scores to be conclusive.

Rater unreliability can be found in all of the above-mentioned errors; the halo effect and the biased grading of learners' work minimise rater reliability by eliminating the equal opportunity for learners to be graded in a fair and objective way. When a rater commits the error of being too severe or too lenient with the scoring, rater reliability cannot be guaranteed. It is possible to still have proof for inter-rater reliability as it is possible for one rater to be equally strict or overly tolerant through many ratings. However, if the results are compared with other raters' results, the tendency of the rater becomes clear and the unreliability becomes apparent. The halo effect, central tendency and a restriction of range leave the rated person with feedback that is not accurate and thus rater reliability is reduced. For instance, if the same person took the same test on two different occasions but deliberately tried to fail one time, the rater who marks the work with a central tendency might still try to push the final scores to the middle. Another factor that influences rater reliability is the above-mentioned criticism on the application of rating scales. Inter-rater reliability could be diminished by the teachers' individual understanding of the descriptors' underlying meaning and lead to various interpretations of the criteria in a scale. Still rater training can increase this type of reliability (Berger 2014: 63).

The overall problem with validity in this context seems to be the question of whether the scores of the writing can even be legitimate when raters do not grade the text in the way they should. For example, when a rater commits the error of central tendency, the question arises whether this rating is significant or meaningful. When a rater is biased and therefore influences the results to favour a student, the task itself might be valid, but the test scores must be questioned. With the rating of students' written compositions, the independence of the criteria in the analytic rating scale must be considered, otherwise, the validity of the scores cannot be given. The same goes for construct validity: when the underlying theoretical constructs of the writing become blurred by the rater and correlations are seen where they should not occur, the validity of the construct is strongly influenced.

Feedback and washback are influenced by rater errors as well. Constructive and useful feedback cannot be received by the learner as the test scores do not reveal their true skills or deficiencies. Although beneficial washback might occur in relation to the students' attitude towards the language or testing itself, the feedback from a poorly rated test does not help the learners improve their language skills. For learners whose work is rated more severely than necessary, washback could negatively influence their motivation or interest in the language.

Based on these theories about the rating process and the connections between rating related issues and their impact on bias, objectivity, fairness and assessment principles, the following study was conducted to gain more knowledge about the application of the CLAAS in the Austrian EFL classroom at upper secondary level.

5. The study

As mentioned in section 3.2., analytic rating scales have been the subject of discussion for a long time, especially in relation to testing and assessment in general, in connection with standardised exams and classroom use. Many scholars have found the analytic scale to be more open to opportunities for differentiated feedback and information on the writer's strengths and weaknesses (e.g. Brown & Abeywickrama 2010; Weigle 2002). Additionally, the fact that the four criteria of the CLAAS are to be treated and graded individually, meaning that one mistake can only lead to the choice of a lower band in one of the criteria, prevents the rater from choosing lower bands for the candidate for one mistake in more than one dimension. This turns the scale into a fairer and more differentiated system for grading written compositions than other grading methods, as discussed in section 3.2.

However, it was also mentioned in section 4, as raters are still human, personal preference, bias and different viewpoints or interpretations of the scale's wording are always part of the scoring process, which leads to the question as to whether raters are indeed influenced by their personal experiences with the L2 and grading, their attitudes towards the scale and their professional experience with scoring and giving feedback in relation to the written work of EFL learners. Lumley's research on the rating criteria and the decision-making process concentrated on the interpretation of the scoring criteria by the raters and issues of first impression scoring, with results showing "a tension between the rules and the intuitive impression, which raters resolve by what is ultimately a somewhat indeterminate process" (Lumley 2002: 1). The intuitive part or the underlying beliefs, perceptions or preferences of the individual rater are what is relevant for this study. Questions as to whether raters can be unbiased, make their decisions not depending on the overall impression they gained from the text or the candidate's performance in other criteria are expected to be answered.

The interest in these issues raters have to face in the process is the foundation of the study along with the curiosity about the impact this area has on the teaching and testing of writing in English as a foreign language.

5.1. Methodology

For the above-mentioned reasons, the following study focuses on a group of teachers who have been working with the scale. As a first step, the participants completed a questionnaire including questions and statements to find out about their experience with assessing writing in general, their teaching experience, their personal attitude towards the CLAAS rating scale and its use, their preferences in relation to the elements of writing and style (e.g. for some teachers, grammar might play a bigger role than layout and organisation or the other way around). The answers were analysed, giving them numbers from 1 (strongly disagree) to 5 (strongly agree) and averages and standard deviations were calculated. As a next step, a review of essays was conducted which these teachers corrected with the B1 or B2 rating scale and in which they gave some kind of written corrections and/or feedback. The teachers were asked to copy two recent test text which they corrected; there was no request for good or bad texts and the teachers could choose which ones to copy. The aim of this review was to find instances of possible influences from the raters' personal attitude on the grading of their students' texts. An example would be when a student has a wide range of vocabulary but lacks grammatical accuracy; in this case, the rater should give high scores for the LSR criterion and only choose a lower band in the LSA criterion of the scale. However, if that rater tends to weight the grammatical accuracy features of a text more than vocabulary range for whatever reason and this is somewhat affected by their personal preferences, attitude or experience, the scoring process might show that the teacher chose a lower band in another criterion (e.g. vocabulary range) to reduce the overall points for the candidate in question. Such an instance can then be compared to the questionnaire the teachers completed, and conclusions can be drawn from their personal preferences in their grading and their feedback on the text. Not only personal beliefs but also the participants' teaching experience and whether they have undergone rater training are related to their way of correcting and applying the rating scale. The guidelines for using the rating scales also play a crucial role as they are the foundation for the teachers' handling of the tool. Knowing the rules of how to apply the scale to a student's text is a basic prerequisite in order to work with the rating scale in an accurate and careful way. As an addition to the above-mentioned factors which could influence the grading process, the rater errors discussed in section 4.3. are also part of the study's focus.

The results of this study then lead into a discussion about the implications of the rating scale for teaching and testing, especially in connection with rater training, fairness, bias and objectivity. Methods of how teachers can maintain a viewpoint that is as objective as possible are proposed and discussed.

Finally, the study's limitations and possible drawbacks are discussed and further research, as well as methods, are suggested.

5.2. Questionnaire

The questionnaire, which can be found in section 9.1., in this study was written to receive information on the participating teachers' opinions, preferences and their use of the analytic rating scales provided by *BIFIE.*

In order to determine the participants' background, the first part asks them to give information on teaching experience and rater training. The second part contains statements which the participants of the study match to their own experience, opinion or practice.

To gain information about the participants, their teaching experience in years was asked in order to compare the participants' career length, which might not only lead to conclusions about the expertise of grading but also give an insight into a possible influence stemming from the fact that teachers who have been teaching for a long time have also experienced more changes in the grading system and therefore have had to use other rating methods, which in turn might influence their grading with the scale in question. The names of the teachers were asked in order to match their questionnaire with the essays they rated. The participants were asked to state whether they have undergone some kind of rater training for writing or speaking in the past as this information also sheds light on the right application of the complex rating scale. For writing and speaking there are similar rating scales and the process of rating is similar in terms of the procedural aspect. It is expected that raters who have undergone training are more efficient users of the analytic rating tool and know how to avoid common rater issues as rater training aims at clarifying the use and practice of these scales.

The participants are reminded of the official terminology and their abbreviations in between the two parts of the questionnaire to make sure that the four main criteria of the CLAAS are clear (task achievement (TA), organisation and layout (OL), lexical and structural range (LSR), lexical and structural accuracy (LSA)).

The second and main part of the questionnaire asks for crucial information in connection with the theoretical framework discussed in the previous chapters of this paper. Thus, the statements are based on rater issues concerning the grading of written compositions with an analytical rating scale, in this case, the two scales used in Austrian upper secondary schools (B1 and B2 level). Issues of severity or leniency, the halo or horns effect and central tendency in the process of correcting students' writing and problems of bias are the foundation for the wording of the statements in the questionnaire. Based on Dörnyei and Csizér (2012: 76-77), a Likert scale was compiled in which participants were asked to rate how much the statements match their own opinion about their grading from *strongly disagree* to *strongly agree*. All of the statements try to elicit information on problematic areas in the grading process. The aim of the survey is to find out how the teachers view their grading and their use and application of the rating scale.

There are six different categories in which the statements can be put and they are based on: first, the *BIFIE* and its guidelines for applying the scale and the *LBVO* (*Leistungsbeurteilungsverordnung*), second, personal preference or opinion, third, the teachers' own experience, fourth the rater issues of grading too leniently or severely, fifth, bias and the halo or horns effect and sixth, central tendency. They can be grouped accordingly, starting with the guidelines. In the questionnaire the teachers received, the statements were not grouped according to these categories but rather put in a random order.

In order to find out whether the teachers in this study are aware of *BIFIE's* guidelines and apply them correctly, i.e. use the scale the way it was designed to, the following statements were written. Additionally, the second statement is concerned with the *LBVO*, as this set of rules must not be ignored in the grading process. The statements were formulated using the official guidelines published by the institute in charge (BIFIE 2014a: 4-6).

- I am aware of the guidelines that *BIFIE* provides for working with the assessment scales.
- I am aware of the laws and guidelines in the *LBVO* concerning the grading of written texts.
- The use of correct but simple structures must lead to the choice of a lower band in the LSR criterion.

- Every mistake is only counted once in each criterion in the correction process.
- I always read a text more than three times.
- With every category, I start with band 6 and then go up or down the bands, considering the candidate's writing.
- When I have an impression of an aspect of the text, I try to find suitable wording in the assessment scale to award a particular band in the relevant criteria.
- The content points (their degree of development) are more important than the other three criteria.
- In my grading process, I consider the first descriptors of each band as more important and thus they weigh more in the decision-making process.

Figure 4. Questionnaire items: BIFIE & LBVO.

Following the statements about the guidelines relating to the scale, the teachers' opinion on the rating tool and its use or personal preference of one or several of the criteria listed in the assessment scale are covered. To gain an insight into that, the following statements were formed which are all based on the descriptors in the CLAAS (BIFIE 2014a: 8-9) and previously discussed issues of fairness and bias (see chapters 5.3. – 5.5.).

- A very important language aspect of writing is grammatical accuracy.
- A very important language aspect of writing is lexical accuracy.
- A very important language aspect of writing is lexical range.
- A very important language aspect of writing is grammatical range.
- Avoiding repetition of vocabulary is highly important for me.
- Avoiding repetition of structures is highly important for me.
- The task must be fully achieved to get a good grade overall.
- In my opinion, *BIFIE's* analytic assessment scales (for level B1/B2) are useful and necessary.
- I find *BIFIE's* assessment scales fair.
- I would rather correct the texts using my own assessment criteria.
- The wording of the scale is open to interpretation.

Figure 5. Questionnaire items: teachers' opinions**.**

Then, questions about the participants' own experience were put in the following statements:

- When I was in school, my own texts were graded using analytic rating scales.
- My texts in school were not graded with an analytic rating scale and I found it rather unfair.
- Grading is a stressful process for me.
- I find it hard to match my impressions of a text with the wording of the scale.

Figure 6. Questionnaire items: grading experience.

The remaining statements are all based on rater issues and try to elicit information on whether the participants show tendencies to grade the writing with these common errors. The basis for these statements was discussed in section 4.3.

The next set of statements was based on the rater issue of bias and the halo effect. These effects and bias in general are not easy to distinguish, as the following statements show. As was already mentioned in sections 4.5. and 4.6., a decision can be biased because of the teacher's error of assigning a lower band in a criterion because the candidate also scored lowly in another one whereas the halo or horns effect causes the teacher to grade on the basis of the general impression of the text instead of seeing it as four different parts which do not influence each other. The following questionnaire items were formulated to gain information on possible occurences of rater errors:

- The use of correct but simple structures must lead to the choice of a lower band in the LSA criterion.
- The use of correct but simple structures must lead to the choice of a lower band in the LSR criterion.
- Missing markers which show the relationship between the candidate's ideas are not an exclusive OL problem but should also result in the choice of a lower band in the LSR criterion.
- Missing markers which show the relationship between the candidate's ideas are not an exclusive OL problem but should also result in the choice of a lower band in the LSA criterion.
- If the word count is not observed, I automatically deduct points in other criteria as well.
- When a candidate scores high in one criterion, I tend to assign more points to others as well.
- When a candidate scores low in a criterion, I tend to deduct more points in other criteria as well.
- If the text is awarded a low band for OL, it cannot receive a good grade overall.
- Ignoring paragraph conventions can lead to the choice of a lower band not only in the OL but also in the TA criterion as one might not be able to follow the thought process.

- As long as the task is fulfilled, the writing does not have to be highly accurate in terms of LSA.
- A candidate cannot receive very high points in one criterion and very few or zero in another one.
- As long as I can justify my rating choices with the scale, I can freely choose a higher or lower band than others might do.

Figure 7. Questionnaire items: rater issues of bias and halo effect.

Information on whether a teacher has a central tendency is elicited in the following statements:

- I tend to end up in the middle section of the bands.
- When in doubt, I rather choose the middle section of the bands in each criterion.

Figure 8. Questionnaire items: central tendency.

Finally, the last two statements try to gain an insight into the leniency or severity of a rater:

- I am a rather lenient rater.
- Whenever I cannot decide between two bands, I choose the higher one.

Figure 9. Questionnaire items: leniency or severity**.**

Together with the written feedback on the graded writing of their students, the teachers' answers provide information on the complex issue of how educators use the scales, whether there are any discrepancies between the way the scales should be used according to the institute which developed the scale, i.e. *BIFIE*, and whether conflicts arise in connection with bias or other previously researched rater issues like the halo effect, central tendency or a rater's tendency to rate more severely or leniently than necessary. The texts, which were written by students in the upper secondary and graded by the participants, include various text types which are among the ones for the written part in the English *Zentralmatura* (BMBWF 2017). The types in these samples are (opinion) essays, reports, informal emails and blog entries.

5.3. Participants

Based on several previous types of research on the issue, e.g. Lumley (2002), seven teachers, who underwent rater training and received instructions for applying the CLAAS to students' writing, were chosen for this study. The seven participants are teachers varying in age and experience but sharing one feature: they are currently all teaching English at the same grammar school in Lower Austria. Their teaching experience ranges from two to twenty-seven years. Additionally, all of them teach English in upper secondary and every one of them claims to have undergone some sort of rater training for writing. The three teachers with the least amount of experience underwent one afternoon of rater training; the other teachers report more training (university courses, seminars, courses). The two teachers with two and three years of experiences had not undergone any rater training for speaking.

To simplify the reading of the analysis and for the sake of clarity, the teachers were referred to by one letter (S, N, P, etc.) and their texts with the same letter and a number (e.g. S1, S2, N1, N2, etc.) in order to discuss them. All questionnaires and texts can be found in the appendix. Additionally, as there are many references made to the CLAAS and its guidelines, they are also provided in the appendix of the paper.

5.4. Results and analysis

To determine the kind of feedback the teachers gave in the questionnaires and their corrections of the writings, the answers, comments and corrections must be analysed and categorised. In order to do that, the statements were grouped and presented in chapter 5.2.

The analysis starts with the participants' answers to the statements in connection with the *LBVO* and *BIFIE's* guidelines as they give some indication of the teachers' general knowledge about how to use and apply the scales. Following this, an analysis of the teachers' personal preferences and opinions is presented and put into relation to their graded student writings. Subsequently, the participants' teaching experience and their answers to statements about their experience are connected to their grading. Finally, information about the previously elaborated rater errors is elicited from the questionnaire answers and then connected to the participants' grading.

5.4.1. Laws and guidelines

The findings in relation to the guidelines and laws provided by *BIFIE* and the *LBVO* are essential for the study and its implications since understanding and applying these rules are fundamental for English teachers in Austria. The scale can only fulfil its purpose of being a tool for improving students' writing skills when it is correctly applied by the raters. The questionnaire asked about information on the teachers' knowledge of *BIFIE's* guidelines and the *LBVO* and their application in eight statements. The statement which resulted in the most interesting findings are discussed in more detail.

First, the survey asked whether the participants are aware of the guidelines provided by *BIFIE* and *LBVO*. All except for teacher J were aware of *BIFIE's* guidelines. Nevertheless, teacher J seems to know many of the guidelines as other statements concerning them were answered as if it were the case. A reason for that could be the participant knowing the rules or most of them but not being aware of where they are from. Another reason could be that the participant learnt about the guidelines through other teachers or instructors and the details of the rules were not discussed or questioned.

The second interesting finding concerns four statements that are related. The two statements "The use of correct but simple structures must lead to the choice of a lower band in the LSA criterion" and "The use of correct but simple structures must lead to the choice of a lower band in the LSR criterion" were given in order to check whether the participants know that they should not count mistakes in more than one criterion and whether they can differentiate between the criteria and what kind of textual features belong to which one, according to the guidelines (BIFIE 2014a: 4-6). In connection to that, the statement "Every mistake is only counted once in each criterion in the correction process" was also part of the survey and almost every teacher knew that, except for one. However, they did not seem to be absolutely sure of that as only one participant chose to answer with *strongly agree*. Interestingly, the participants were all sure that the development of the content point in a text is not more important than the other criteria. Four out of the seven teachers knew that correct but simple structures lead to the choice of a lower band in the LSR criterion, but only three were sure that it is not a problem concerning the LSA criterion.

Although almost all of the teachers stated that they knew *BIFIE*'s guidelines, only two teachers said that they read the texts more than three times as recommended by the instructions (BIFIE 2014a: 4-6). This is a significant finding for this research and raises the important question as to whether the teachers can mark the writing in an analytic way when they do not read the texts four times. Clearly the four criteria which all deal with different aspects of the text must be treated separately and thus it is necessary to read the text four times. Reading the text once for each criterion, concentrating on the relevant descriptors of the textual feature does make sense in order not to confuse, for example, an accuracy issue with a range issue in the writing. As mentioned in the discussion of the different rating scales in section 3.2., the criteria must be treated separately to maintain the analytic character of the scale (Bachman & Palmer 1996: 211). This finding suggests that the teachers follow a rather holistic approach and try to match their initial thoughts or first impressions of a text with the scale, as mentioned in the definition of holistic rating scales by Brown and Abeywickrama (2010: 283). To investigate this issue, more research would have to be done and think-aloud protocols would be helpful, but this observation suggests that the participants might not use a rating approach which clearly distinguishes between the assessment criteria.

An indication that the teachers are aware of *BIFIE*'s guidelines is that six out of seven follow the instruction of starting the grading process in each criterion with band 6 (BIFIE 2014a: 4-6). Although these guidelines also state that the descriptors in each criterion are arranged in an order that lists the most important ones first and continues in descending order, only one participant states to include this in the grading process. The statement asking about the impression of the text might be problematic as it was intended to find out whether an initial feeling or idea of the writing influences the analytic approach the teachers should use. However, using the word impression does not fully express this idea and the statement can be interpreted in various ways, including the understanding that the impression refers to the observations the teachers made in each criterion and thus including the analytical grading approach. As this became only apparent after the questionnaires were collected, the statement's wording could not be changed into something clearer.

After looking at the texts the participants corrected, two of the texts stand out in the way they were marked. On the one hand, there are two texts, provided by teacher T, which were corrected using the same abbreviations as in the assessment scale and guidelines, showing which mistakes were counted in which criterion. On the other hand, there are two texts, provided by teacher R, which were not corrected with the help of the assessment scale; the teacher used three assessment criteria which were given to the students as a part of the prompt. The texts were then corrected by writing down the corrections or improvements in the text, not indicating what type of error occurred. It is also hard to reconstruct the grading or thought processes of the rater and how the final points were allocated. The other five teachers marked the texts using the proper assessment scales but they all used their own abbreviations to indicate whether the mistakes in the texts are grammatical or lexical ones. Mostly, they did not go beyond these language aspects and only a few comments hint at organisational or task fulfilment concerns. Another interesting observation was that none of the participants used the suggested colour coding in the grading process. A commented writing performance was published on the official website of the *Zentralmatura* by the Ministry of Education in which the use of different colours for categorising the errors in the various criteria is recommended (Bundesministerium für Bildung 2013). This sample correction also illustrates the clear distinction between the four rating criteria, providing four copies of the same student essay, each of which illustrates the corrections of one criterion at a time (Bundesministerium für Bildung 2013).

Altogether, the participants stated that they were familiar with the guidelines set by *BIFIE* to grade written compositions and the laws of the *LBVO*. Yet no one was absolutely sure of these rules and the teachers' answers in relation to them were tentative or contradicted the guidelines. Only one teacher marked the text using *BIFIE's* abbreviations; the same teacher also knew most of the rules according to the questionnaire.

5.4.2. Personal preferences and opinions

In this section, the participants' answers to statements are analysed in connection with their opinions on elements of the criteria and their importance in relation to the other criteria in the assessment scale. The teachers' answers are compared and discussed and examples in the graded texts are given to show the connections between the teachers' opinions and their actual written feedback on their students' exam texts.

First, the statements asking about the participants' personal preference for a specific textual feature in the scale display their priorities. Three out of the seven teachers weighted the language aspects of the scales equally, just like the CLAAS intended them to be. Interestingly, the three still chose the answer *agree* for these statements about the importance of certain features throughout the questionnaire and no one agreed strongly, which might be connected to their not wanting to choose the extreme end of the Likert scale as there is no feature for them that is more important than the other ones. The other four participants vary in their answers, but a tendency can be identified towards the features regarding lexis. The average answer to the statements was 3.6 and 3.7 about structural or grammatical features and to statements about lexical features, it was 4.3, showing a general tendency for the raters' placing more importance on them. The statement with the lowest average number is the importance of grammatical accuracy. This is a rather unexpected finding as this textual feature was the decisive element of texts written in the EFL classroom for a long time as writing was often graded by deducting points for accuracy mistakes and many teachers put the greatest emphasis on this language feature.

An illustration of these opinions can be found in one participant's grading, teacher S. The teacher agreed that structural range is very important and even strongly agreed that lexical range is very important in a text. In text S1 there are numerous accuracy issues, but the LSR criterion is also not on a high level. Teacher S assigned band 6 for the LSA criterion and band 5 for lexical and structural range. On the students' language level, in this case, B1, the range could be wider, but band 6 could also be allocated to this student. This, connected to the answers in the questionnaire, could hint at the teacher putting more emphasis on range than accuracy and thus teacher S chose band 5 instead of 6. This is a particularly interesting observation as the pass mark for each criterion is 6 and allocating band 5 results in this criterion being marked as negative, meaning that the descriptors in band 6 do not apply. Text S2 shows indications of the same kind. Although the student makes numerous accuracy mistakes and shows as many deficiencies in the LSA criterion as in the LSR criterion, if not more, teacher S chose to allocate band 6 for accuracy but only band 4 for range. In both cases, teacher S could have easily chosen a higher band in the range criterion; what stands out is this choice of assigning lower bands for the limited range, especially since the students both expressed themselves in a way that the reader clearly understands what is meant and tried to vary their expressions and structures with only some restrictions. Whether teacher S chose the lower band because of their personal opinion on the importance of range or because of other reasons cannot be discussed here. Nevertheless, the information available shows such indications and conclusions can be drawn from that.

Teacher J provided two graded texts from students who attended the 7th grade in upper secondary. Teacher J stated that lexical features like avoiding repetition or lexical range and the structural issue of repeating forms are more important than accuracy issues in a text. Looking at the texts J1 and J2, there is no indication of teacher J's personal preference of one criterion over the others. It is obvious that this participant places great importance on lexical issues from the answers given in the questionnaire, but there is no evidence that might suggest special attention to one of the features the teacher claimed to be more important in the actual grading of the two texts. Generally, the grading appears to be thorough, weighing the four criteria equally.

Another finding is related to the texts which were not corrected with *BIFIE*'s assessment scales. Teacher R used other criteria; whether they are their own criteria or taken from previous rating scales in use remains unclear. These criteria seem somewhat minimalistic as there are not many descriptors and the grading process is not transparent as it is only known that there are six points to score in each criterion but not how the teacher decides how many points to allocate in each one. Interestingly, teacher R argued to place great value on lexical range, but the assessment criteria used for grading the texts do not have an extra criterion for this textual feature. The scale includes lexical range in a criterion named *Ausdruck* (which is *expression* in English), but this criterion also deals with spelling and comprehensibility, among others. The assessment scale, i.e. the CLAAS, which should be used in upper secondary, deals with lexical range in more detail than these assessment criteria and might, therefore, be more suitable to the teacher's preferences. Apart from that, grading texts with the CLAAS is more transparent and prepares students for the writing part in the *Zentralmatura*. Furthermore, due to the many descriptors in the scale, the students receive much more feedback on their strengths and weaknesses than the criteria used in this example of teacher R's grading. These advantages of analytic scales have already been discussed in connection to the different rating scales in section 3.2. and emphasised by Brown and Abeywickrama (2010: 285). Moreover, as already mentioned before, being graded with the CLAAS in the upper secondary prepares the students for the requirements in their school-leaving exam. Hence the use of this scale is recommended and preferable to other assessment scales.

Teacher S, J and R were the participants who did not weight the textual features as being equally important. The fourth participant who did not do that, teacher A, provided two graded opinion essays written by students in the 7th grade on a B2 level. In the questionnaire, teacher A said to find avoiding repetition of vocabulary and structures as well as having a wide lexical range as being very important. While both texts are on a high language level, text A1 has examples of repetitive vocabulary use in it. The student uses the word *gap year* six times and never replaces it with a synonym or paraphrases, which he or she should be able to do on a B2 level. Although teacher A claims to place great importance on avoiding such cases of repetition, the chosen band 9 is high. This score in the LSR criterion could be influenced by the student's good use of a variety of structures or their broad range of vocabulary, but as the lexical range is not very wide either, the decision to allocate band 9 is hard to comprehend, especially in comparison to text A2. The lexical range in the second graded text, A2, is much higher and the student varies formulations frequently. Additionally, the words and phrases are more suitable to the task, there are not many mistakes which could be related to this criterion and the overall language level seems to be higher. Still, text A2 received the same score on the LSR criterion as text A1 did. Generally, the grading of teacher A does not seem to be influenced by their personal opinions or preference of one criterion or language aspect, but it can also not be fully comprehended which choices were made for which reasons.

To fully analyse the connections between the opinions, the comments and feedback of the teachers on the writing, the three participants who stated that they find all features equally important must be analysed as well. Teachers T, P and N agreed that accuracy, range and avoiding repetition of structures and vocabulary are similarly significant in a text. Indications of mistakes or correctness in the texts in all four criteria were expected findings in their corrections. Various observations could be made in their graded texts. Starting with teacher T, who generally applied the CLAAS and its guidelines the way it should be done more than the other teachers, also seems to follow his statements made in the questionnaire. In the two texts, one aspect can be highlighted: teacher T is rather precise in his corrections and always indicates which mistake belongs to which criterion, but it is not absolutely clear how the teacher comes up with the final scores for task achievement and organisation and layout. The teacher sometimes indicates that there is an organisation or task fulfilment problem, but it is not easy to make sense of what kind of issue it is and how the student could correct it. Obviously, teacher T could talk about these issues with the student in more detail; it is also often easier to mark accuracy and range issues than issues related to task achievement and organisation and layout for the simple reason that the latter are often more complex and difficult to comment on in a few words. Nevertheless, as teacher T marks the texts thoroughly, the expectations of finding more detailed comments on all four criteria are rather high. Finding that there are only a few comments with unclear references could hint at a prioritisation of the two language-related criteria on the assessment scale. Additionally, sometimes teacher T writes two abbreviations next to one mistake, indicating the scoring of one mistake in two different criteria. This is a rather surprising discovery since this teacher seemed to know the guidelines and application of the scale very well from his answers in the questionnaire. Follow-up questions for this teacher would include asking about how the mistakes are considered in cases where the teacher writes, for example, TA/LSA or OL/LSR next to a sentence with an underlined word or phrase. Similar observations could be made in the texts graded by teacher N. Although this participant argued that she weights the importance of the four criteria equally, her choices for low bands in the TA and OL criteria are difficult to reconstruct. The two texts are marked in terms of the two language criteria, but there are no indications of why the teacher chose the bands for the other criteria. Especially in text N2, in which the student only scored band 3 in task achievement and band 2 in organisation and layout, the written feedback or corrections from teacher N do not provide explanations about the final points.

In contrast, teacher P, who is the third one of the participants who weighed the textual features equally in the questionnaire, continuously used abbreviations which make the decisions of the final scores more comprehensible. In comparison to the other two teachers, T and N, teacher P highlights parts of the texts and adds comments like “cohesion”, “TF” (task fulfilment) or inserts signs which indicate that something was missing in the text or layout. The written feedback of teachers T and N might be more influenced by their subjective views of the textual features and place more importance on the language criteria without being aware of it. However, their scoring of the TA and OL criteria does not seem to be influenced by that as the lack of comments does not result in higher points in these criteria. Thus, it cannot be assumed that their preference for the language criteria affects their scoring, but it can definitely be observed that it affects their commentary on the texts. Teacher P's corrections and comments in the texts mostly reflect her answers in the questionnaire and therefore it can be assumed that the weighing of the criteria is not influenced by personal preferences in the case of this participant.

Among the survey items relating to opinions and preferences, the statement "The task must be fully achieved to get a good grade overall" was mostly agreed upon by the teachers, with an average of 3.8 (an overall tendency to agree) and a standard deviation of only 0.6. As task achievement is one of the four criteria in the analytic assessment scale, it makes sense to attach importance to it. However, the task does not have to be **fully** achieved in order to get a good grade overall as long as this criterion is partly completed and the other criteria are not too low. TA is concerned with other issues in the descriptors as well. The use of the word fully could explain why the average number is not higher as this statement is only partly true. The graded texts also show that the participants did not insist on full task achievement as a requirement for an overall good grade, as there are some gaps in this criterion which only led to a lower band in the TA criterion. In this context, an example is provided by teacher S, who agreed strongly with this statement. The teacher claims that the full achievement of the writing task is necessary to receive a good grade on the text. Looking at the graded exam text, it becomes obvious that this claim does not mirror the grading of this teacher. The sample text does not entirely fulfil the descriptors in the task achievement criterion and thus the teacher allocated band 6. If the other three criteria had each received band 10, the text could have got a *Gut* (=second best grade) overall. In this case, lower bands were also assigned in the other criteria, so the text did not receive bands that would result in a good grade. Consequently, either the teacher answered the statements in the questionnaire one way and grades the other way or their opinion about task achievement influences their decision-making process in the other criteria. After looking at the other three criteria and the points allocated to them, the case does not become clearer as there are various reasons to deduct points in each criterion. Whether the reasons for this discrepancy between the questionnaire answer and the grading are of the one or the other kind is hard to say and cannot be definitely ascertained in this context.

In addition, there are three statements in the questionnaire which were chosen to find out about the participants' attitude towards the assessment scale itself, its usefulness and fairness. The question of whether the teachers would like to use their own assessment criteria instead was posed for the same reason. The idea was to see if there are connections between teachers who have a rather negative view of working with the scales and their knowledge of the guidelines and opinions of the importance of different textual features. In this research, the teachers seem to share the belief that the scales are useful, fair and none of the participants would like to use their own criteria for grading texts. However, they all acknowledge that the wording of the scale is open to interpretation.

In conclusion, the insights gained about the impact of the participants' views on their grading do not show a general tendency or homogenous results. The slight tendency to place more importance on lexical aspects in a text became evident and a case of grading affected by this preference could be illustrated by teacher S, but at the same time, two other participants' ratings were not affected by their views. The most noticeable observation was made in the grading of teacher R, who did not use the recommended rating scale and instead graded the exam texts with another condensed scale, which only had three assessment criteria and descriptors which were substantially superficial and reduced to the absolute minimum. Considering the great value teacher R placed on the lexical aspects of a text, the descriptors connected to lexical issues in this scale were correspondingly limited. Further observations could be made in the comparison of the questionnaire answers and corrections of the three teachers who seemingly did not prefer one criterion over another. In two of the three teachers' marking, the tendency towards language criteria became evident in terms of the frequency of comments or corrections. Yet there was no indication of this affecting their scoring decisions. The seven participants mostly agreed on the necessity of the task being fully achieved to receive a good grade, although the majority did not assign lower bands in other criteria even if there were errors or gaps in this regard. Finally, the teachers all share a positive opinion about using the analytic assessment scales provided by *BIFIE*.

5.4.3. Teaching experience

Following the personal attitude of the teachers, this section deals with insights about the participants' experience of teaching and therefore grading. This information is analysed and discussed, but a connection between the statements about experience and the graded texts does not give enough insights. More research in the form of interviews would shed more light on the question as to whether a teacher's experience with rating scales influences their attitude or their own grading process positively or negatively. Nevertheless, conclusions can be drawn from the answers in the questionnaire to provide some information on the connection between the teachers' experience and their attitude towards the scales and the grading process.

One striking finding is the fact that none of the participants' texts were graded with the help of analytic rating scales when they were pupils. For the more experienced participants, this was an expected outcome as the *Zentralmatura* and *BIFIE's* analytic rating scales were only introduced to the EFL classroom in Austria in 2014 (Lopatina 2016: 7; BIFIE 2014a; BIFIE 2014b). For the newer teachers, it is a rather surprising result, as the youngest participants are only 26 and 27 years old and analytic rating scales have been in use at the university for some time and it can be assumed that they learnt about them in the course of their studies. Four out of the seven teachers found being graded without an analytic rating scale unfair, which indicates a positive attitude towards the use of the CLAAS. Together with their agreement with the statements on the usefulness and fairness of *BIFIE*'s assessment scales, it can be concluded that the overall attitude of the teachers in this study is quite positive.

Another interesting finding in connection to their rating experience is that three teachers stated that they find it hard to match their impressions of a text with the wording of the descriptors in the scale. The same three teachers also stated that they find grading to be a stressful process. This finding suggests that the difficulties the teachers have with applying the scale to a student's writing influences the stress level of the raters. There are undoubtedly numerous other factors that contribute to a teacher's stress level in the grading process, but this observation is a noticeable connection between the two statements and answers in the questionnaire.

Overall, the statements about the teachers' experiences show their positive attitude towards using this analytic rating tool.

5.4.4. Rater errors

Following the analysis of the teachers' statements about their experience of teaching or grading, this section aims at showing what kind of hints about rater errors could be found in the analysis of the questionnaires and the texts graded by the teachers. It is clear that this analysis cannot show unambiguous or definite results which can give a diagnosis of issues the raters have. It can only present the findings from two graded texts per teacher and does not try to make any universally true assertions. Moreover, some rater issues cannot be distinguished easily, for example, it is hard to say why a rater decides to assign higher or lower bands to a criterion even though the text should clearly be in another band. Similarly, when a rater chooses to assign lower bands in two criteria for the same mistake, it cannot be said whether this is done because the rater tried to match their general or first impression to the scale or because the rater is biased due to other reasons. To make statements about definite issues and their reasons behind them, far more research would have to be done, for example, think-aloud protocols, interviews or similar, in order to gain more insight into the raters' grading process (e.g. Barkaoui 2011; Cumming, Kantor & Powers 2002). Nevertheless, some findings can be categorised and analysed, bearing in mind that these are no definite diagnoses.

First of all, the participants' answers in the questionnaires show some tendencies for their grading to be similar to rating issues discussed in section 5.3. The statements which were formulated to find out whether there are instances of the halo or horns effect or biased decisions resulted in insightful findings. The statement claiming that the use of correct but simple structures must lead to the choice of a lower band in the LSA criterion, which was already discussed in connection with the scale's guidelines, showed that three participants agreed that this should be done in the grading of a text with the scale. Interestingly, two of them also agreed with the similar statement that includes the correct version of choosing a lower band not in LSA but in LSR criterion; the third person neither agreed nor disagreed on that. This suggests that teachers T and R would, therefore, choose a lower band in two criteria for the same mistake or they simply do not know which criterion deals with this type of mistake. In teacher R's corrections, it is not possible to reconstruct the grading process as the CLAAS was not used and the teacher does not indicate what kind of mistakes the student made. In contrast, teacher T's corrections are rather transparent and due to the use of the CLAAS abbreviations, his classifications of the students' mistakes are comprehensible. In connection to the two language criteria, the teacher clearly indicates which mistake is counted in which criterion and there are only a few cases in which one identified mistake is categorised as being both an LSR and LSA issue; that happens only in cases where the student used a word that does not fulfil the purpose of what they tried to express and at the same time the wrong tense or word form was used. In such a case, there are indeed two mistakes and teacher T correctly sees two separate problems, one with range, and one with accuracy. However, the corrections of teacher T in the texts and the answers in the questionnaire are inconsistent and what becomes apparent is that this participant might have agreed to rate the range issue in the accuracy criterion by mistake as hr knows the guidelines very well and also applies the descriptors accordingly and consistently.

Teacher R and his consistently deviating answers are a noteworthy finding in the participants' answers to the questionnaire. In many cases, e.g. the statement in which the claim was made that language accuracy is not as important when the task is achieved, teacher R's answers differ from those of his colleagues. While this often shows that this participant is not as familiar with the rating scale and its guidelines, it is also not surprising since he is also the only one who did not use the CLAAS to correct the exam texts. Therefore, his answers suggest that he did not only not use the assessment scale to correct these texts, but that he generally does not use the scales provided by *BIFIE*.

The next two statements in connection with the halo or horns effect or bias in general asked the participants whether they agree that missing markers which show the relationship between the candidate's ideas are not an exclusive OL problem but should also result in the choice of a lower band in the LSR or LSA criterion. Both statements are problematic since they indicate counting one type of error in two different criteria, thus failing to treat them separately and to use the scale analytically. Out of the seven teachers, two argued to choose a lower band in the OL and LSA criteria and one teacher said to choose a lower band in the OL and LSR criteria. As the statements address a coherence issue connected to the choice and variety of words or phrases students use in their text, it was expected to find results which tend to assign the additional lower band in the LSR criterion. This criterion deals with the use of words and varying formulations among others, making it easier to categorise the absence of linking devices as an issue in the LSR criterion than in the LSA criterion. Therefore, it is rather surprising that two teachers would choose a lower band in the accuracy criterion and the organisation criterion as the missing markers issue might be a related range issue but it is definitely not an accuracy issue. Looking at the graded texts of the three teachers, there are not enough comments in the margins of the texts to determine whether such errors were counted in two criteria or not. Nevertheless, the participants' answers indicate that the above-mentioned horns effect could affect their decision-making process as they are influenced by a negative impression in one criterion and transfer this notion onto another assessment criterion (Belludi 2010).

A similar issue is addressed in the statement "If the word count is not observed, I automatically deduct points in other criteria as well". Although the word count is an essential part of TA, there is room for varying the length of the writing and a text in which the word count is not observed can still only result in the choice of a lower band in the that criterion. Five out of the seven participants disagreed with the statement above and would not let the word count influence other assessment criteria. Of the two others, one teacher neither agreed nor disagreed and the other is said to assign lower bands in other criteria as well. This participant, teacher S, provided two graded texts in which the students both observed the set word length; thus, there is no example to obtain proof for teacher S' statement. Still, this claim in the questionnaire suggests that teacher S' grading is influenced by the negative outcome in one criterion or even one descriptor in a criterion and transfers this impression onto other, usually separate criteria, which, as discussed above, is an example of the horns effect (Belludi 2010).

In connection with a similar issue, the two statements "When a candidate scores high/low in one criterion, I tend to assign/deduct more points to others as well" mirror each other to find out whether a halo or horns effect occurs in the grading process. In the questionnaires, these two statements were mostly disagreed with, except for teacher A saying that she neither agrees nor disagrees with deducting points when the student scores low in one criterion and teacher R agreeing that he assigns more points for a text in which the score was high for one criterion. As teacher A's answer does not suggest that she is affected by the points in other criteria, this answer is, if anything, only a sign of uncertainty as to how the scale is used or how to apply the provided guidelines. In contrast, teacher R's claim suggests that a halo effect could influence his grading process. However, since teacher R did not use the CLAAS to correct their students' essays, a comparison of his statements in the questionnaire and the texts cannot give more insights into this matter.

The next statement in the questionnaire also aims at eliciting information on whether there are any discrepancies between the distinct assessment criteria and the participants' application of them. The sentence "If the text is awarded a low band for OL because it is confusing, it cannot receive a good grade overall" again asks for the reverse halo effect, i.e. the horns effect. As incoherent or confusing writing makes the reading of a text difficult, the raters might be influenced in their grading of other textual features. If the confusion only stems from an organisational problem and the language used does not contribute to the problem, the rater must not be influenced in their choices of bands in the other criteria. Three of the participants agreed with this statement and only one disagreed. On the one hand, this could suggest that the three teachers are influenced by the horns effect as the one negative textual feature affects their perception of the other criteria (Belludi 2010). On the other hand, the confusion might be also caused by problems in connection with the descriptors in the task achievement criterion. In this case, it is indeed hard for the student to receive a good overall grade when the confusion is the result of issues in two out of the four criteria. Text N1 is an example of a text that received a low OL score. This essay was awarded band 2 for OL and also only received band 3 in the TA criterion, band 1 in LSR and band 2 in LSA. The text is confusing and lacks most of the features described in the OL criterion. Moreover, teacher N also agreed with the statement in question. However, the confusion of the text does not seem to be the decisive factor why the other criteria were also given low bands since the text shows many problems in all four assessment criteria. Thus, it cannot be clearly determined whether the confusion led to the choice of lower bands in other criteria.

A related issue is addressed in the next statement, in which the participants were asked whether ignoring paragraph conventions can lead to the choice of a lower band not only in the OL but also in the TA criterion as it might be hard to follow the thought process. Although this statement seems reasonable, the two criteria should again be treated separately and an issue like paragraph conventions is clearly categorised as an organisational matter by the CLAAS (BIFIE 2014a: 8-9). Nevertheless, the majority of the participants agreed with the statement and thus acknowledge that an error in one area could influence their assessment of another, usually distinct area. Similar to the example above, this suggests the occurrence of the horns effect (Belludi 2010). Again, examples like text N1 show that it cannot be established whether the rater chose a lower band in the TA criterion because of the negative impression of the OL criterion or because there are too many other problems with the text. In this example, there are enough reasons to assume that teacher N was not influenced by the OL criterion in the decision of a low band in the TA criterion as there are numerous gaps and errors in both criteria to award low bands.

The next statement in this category of rater errors is “As long as the task is fulfilled, the writing does not have to be highly accurate in terms of LSA”. Only one participant, teacher R, agreed with this claim, which indicates that the rating of one criterion is again influenced by the rating of another. In this case, the positive impression of task achievement could have an impact on severity or leniency in the rating of the language accuracy criterion. When the assessment scale is used in such a way, the rating shows signs of the halo effect (see section 5.3.2.). As teacher R is the only one who agreed to this statement and also the only one who did not use the CLAAS to correct his students’ texts, there is no evidence to support this claim.

Another item in the questionnaire connected to the halo or horns effect is more general and says that a candidate cannot receive very high points in one criterion and very few or zero in another one. As already mentioned in section 3.4., the four assessment criteria are to be treated separately and thus it is indeed possible for a student to score high in one area and low in another one, although it might be uncommon (BIFIE 2014: 4-6). In the questionnaire, four teachers agreed with this statement, which indicates a tendency towards basing their decisions about one criterion on the choice of band they made in another one, i.e. a student’s performance in one area is assessed by not only looking at the accomplishments in this area but also by including the student’s accomplishments in another area. When these influences lead to a more positive result, the halo effect can be observed; when they lead to a more negative result, the horns effect is indicated (see section 5.3.2.). In the graded texts there is no instance of a result which shows a very high and a very low score in the assessment criteria. Most scores are close together with the majority being only 1 or 2 bands apart. In text T1 the widest gap between two bands can be found with an LSA score of 2 and a TA and OL score of 6. Therefore, there are no examples for this statement which, on the one hand, could indicate that the teachers’ scoring of a criterion is indeed influenced by the performance of their students in other criteria or the rater error of restriction of range occurs. On the other hand, it could also indicate that the students’ writing skills are on similar levels throughout all four criteria. Looking at the students’ writings, the latter seems more plausible as most scores can be explained with the respective band they received in each criterion.

The last statement which deals with the halo or horns effect or bias is “As long as I can justify my rating choices with the scale, I can freely choose a higher or lower band than others might do”. Although the wording of this statement in the questionnaire could have been clearer, this sentence suggests that raters can choose any band in the scale as long as they can find a way to explain it with the scale. This approach is not how the scale is supposed to be used and indicates a general impression grading which is then presented as an analytic procedure (Engelhard 1994: 99). Five teachers neither agreed nor disagreed with this claim and only two agreed, which is not a surprising result as the teachers generally seem to know how to use the analytic scale and how it differs from a holistic one. The two teachers who agreed with the statement could be biased by their general impression of a text in their application of the assessment scale. However, the two participants are teacher T and J, who more or less consistently showed that they are well aware of the guidelines and their corrections in the texts also showed that they know how to apply the scale. Additionally, there is little indication that their grading process is a holistic one as both teachers' graded texts are marked with a coherent and rather transparent system which demonstrates the analytic character of the CLAAS. Consequently, the statement might have been interpreted in another way than it was meant and the two teachers might have thought that they stated that assigning different bands than other teachers do is fine as long as they base their choices on the scale. In the process of analysing this part of the questionnaire, the ambiguity of this statement became apparent and can thus not be considered a significant insight.

To find out whether the participants' grading shows any signs of central tendency, two sentences were given to agree or disagree with: "I tend to end up in the middle section of the bands" and "When in doubt, I rather choose the middle section of the bands in each criterion". Teacher J agreed with the former statement and nobody agreed with the latter one. The two texts provided by teacher J do not show evidence of central tendency. Although text J2 was rated 6-8-6-6, the writing seems to actually fit this assessment. Text J1 was rated 7-9-7-5, scores that are not too close to the middle. Actual indications of central tendency would have to be scores which are consistently around the middle section of the bands, i.e. band 5 or 6 throughout the rating (Engelhard 1994: 99). As teacher J varies the score, there is no evidence of central tendency. Yet these two texts do not give enough information on the teachers' grading habits and thus a judgment cannot be made about whether or not teacher J's grading is affected by central tendency.

As discussed in section 4.3.1., grading a text too leniently or severely is also classified as a rater error (Saal, Downey & Lahey 1980: 417). In the questionnaire, the participants were asked whether they think that they are rather lenient raters and whether they choose the higher band whenever they cannot decide between two bands. None of the teachers would describe themselves as lenient raters, four of them neither agreed nor disagreed with the statement. Teacher J even strongly disagreed with the statement, implying that the opposite is the case and that this teacher is a rather severe rater. This matches teacher J's answer to the second statement, with which she disagreed. The other participants also either disagreed or chose the neutral answers, except for teacher N, who agreed with choosing the higher band in case of doubt. Looking at teacher J's graded texts, it is not easy to find out whether the scores are higher or lower than the student deserves. In text J1, it could be argued that the accuracy score is too low and band 6 could have been chosen as the descriptors of this band match. Otherwise, the scores of both texts are mostly comprehensible and seem to reflect the students' writing skills. Teacher N's claim of choosing the higher band in case of doubt is also not easy to prove with the two provided texts. Overall, the chosen bands make sense and seem to represent the students' skills accurately. The only choice which shows that teacher N may not have chosen the higher band is in text N1, in the LSR criterion. Teacher N chose band 1 but band 2 applies to the writing as well. However, this is only one example and the teacher might not have been in doubt over this choice; it is not a very informative observation as it is only one instance which can be argued about.

In conclusion, the findings in connection to rater errors do not show clear tendencies. Although the participants at times answered the questions in a way that would suggest bias or other rater errors, the analysis of their corrected student texts often did not confirm this supposition or the findings were inconclusive. The participants' answers mostly hinted at a biased grading approach in terms of the horns effect, which could only be seen a few cases in the texts. Altogether, the graded texts showed that the teachers do not seem to be highly influenced by any rater errors with only a few exceptions. Whether these exceptions are indeed instances of rater errors remains obscure as more research, especially looking at more essays of each teacher, would have to be done in order to confirm this assumption. Although insights could be gained into the rating of this teacher group, the study shows some substantial limitations which will be discussed in the following section.

5.5. Limitations of the study

In the course of analysing this research, a few significant limitations became apparent. First, the findings of the questionnaires and the written feedback in the graded texts only provide limited insights into the actual decision-making process of the teachers. As the participants gave their answers to the statements separately from doing any actual rating, their answers might not reflect what they actually do and how they form their decisions in the real rating process. Without further research, for example, using think-aloud protocols, the teachers' rating approaches can only be reconstructed in a restricted way.

Second, the wording of the scale and the teachers' interpretation of the descriptors and also the guidelines play an equally important role. In many cases, it is not possible to say whether an assessment choice was made because of an individual interpretation of the scale or because of the occurrence of a rater error. This research can only present observations made on the basis of a limited number of graded essays and the participants' information on their use of the rating tools. Additionally, the questionnaire did not include further questions about the wording or interpretation of the rating scale and only broached the topic. This is why many statements and findings in the study could also be caused by the teachers' individual interpretation of the rating scale and its descriptors.

Third, the study included a small number of participants and also a small number of student writings. In order to find a teacher's rating pattern and investigate what kind of errors or bias might or might not occur, a larger number of essays graded by the same rater would have to be analysed. This would give a better understanding of the teacher's rating process, especially in combination with interviews, think- or write-aloud protocols or similar.

6. Discussion and implications for teaching and assessment in the EFL context

Conclusions can be drawn from the insights gained by this research and the conducted study. They have considerable implications for teaching and testing. They do not necessarily contain new information but can be related to findings from previous research on the subject.

As mentioned in section 1., many methods have been used to obtain more information on the decision-making process of teachers, but it is not possible to reconstruct the exact relationship between a rater, the text and the assessment scale (Lumley 2002: 246). The study in this paper confirms that the process is difficult to reconstruct and only assumptions can be made about why a rater chooses to grade a textual feature in one way or another. Whether or not the assumptions made in this work reflect actual rater issues or bias, they still show that there are some discrepancies in the rating process of the participants on the one hand. The study also shows that the teachers use the scale in a rather correct and fair manner and mostly understand how to apply the descriptors of the scale on the other hand. For teaching and testing in the EFL classroom in Austria, these findings again show how important rater training is and that a fair use of the analytic rating tool can only happen if the teachers are aware of its rules and are trained in its application to students' writing.

The consequences of the findings are complex; first of all, the correct application of the CLAAS is meant to result in the use of the scale as a feedback and washback tool, as Hughes describes, especially when they are made available to the learners (2003: 105-6). Washback, which was described in section 3.1. as an assessment principle that allows learners to benefit from the formative feedback they receive during their learning process, cannot be guaranteed when the scale itself is not applied in a correct or fair way or when the feedback lacks information on one or more of the assessment criteria (Brown & Abeywickrama 2010: 38). In connection with the study, this means that some students might not have gained enough information on their current language skills and the areas in which they have to invest more time and work to reach a higher language level. If no additional feedback is provided, the students might not know how to review their work in order to learn from their mistakes and further develop their skills, which mostly occurred in the TA and OL criteria in this study. Secondly, intra- and inter-rater reliability are influenced by a biased or wrong application of a rating scale. As already mentioned in section 3.1., the consistency a rater should have in order to guarantee high intra-rater reliability and the avoidance of bias and rater errors to enhance inter-rater reliability can both be increased by rater training (Brown & Abeywickrama 2010:28). This means that the teachers in this study who did not differentiate between the four assessment criteria or whose grading showed tendencies towards a biased approach could improve their rating by attending more rater training. However, as previously discussed, the nature of the scale has a strong influence as well and interpreting the wording of the scale one or another way can also affect the rating process immensely (Berger 2014: 63). It can therefore not be exactly determined how raters are influenced in their grading and where their bias stems from. Berger also proposes rater training as an attempt to approach this problem as raters can be taught how to interpret the scales accordingly to prevent low inter-rater reliability; there is still no guarantee as problems with the scale can also originate from the nature of the scales (2014: 63).

Another interesting connection can be drawn between the results of the study and Lumley's study which was mentioned in section 1. Lumley concentrated on the connection between the wording of the scale and the raters' application of it and concluded that the link between the descriptors of a scale and a text's features cannot definitely be identified (2002: 246). His theories about the reasons for that can also be applied to this study. His research confirmed that the use of a new rating scale is often used as a tool to justify a rater's individual impressions, which could definitely also be the case in the rating of the teachers in this study (Lumley 2002: 267). The teacher is still the one who decides what is most important in a text in the framework of the rating scale (Lumley 2002: 267). This paper's study has shown that the teachers are indeed able to use the rating scale in such a way and put more emphasis on certain areas of the analytic rating scale depending on their own notion of what is most essential in a writing piece. Additionally, the fact that the teachers use many different abbreviations to indicate mistakes or improvement suggestions made it difficult to reconstruct the grading process and the classification of mistakes. Not using the same abbreviations that are used in the assessment scales and their guidelines could also indicate that the teachers did not adjust their rating approach to the new circumstances but kept their previous rating style, as suggested by Lumley (2002: 267). For teaching and testing, this again poses the risk of low inter-rater reliability and a lack of feedback and washback since the learners can still only receive detailed commentary on areas which their teachers prefer.

In all of these complex issues two main implications for teaching and assessing arise: firstly, a transparent and comprehensible use of the rating scale can improve the scale's value as a feedback tool and increase washback and secondly, rater training can increase reliability and eliminate many rating related issues. Although the participants in this study all underwent some kind of rater training, the necessity of continuous training becomes evident. As long as humans correct the writing of other humans, there will always be rater issues or other discrepancies in the rating process. However, the problems can be minimised by constantly working on a common interpretation and methods for applying the scale. While Lumley (2002: 267) argues that rater training often only leads to a new way to justify the teachers' general impression of a text, other scholars emphasise its importance for increasing reliability and objectivity (e.g. Brown & Abeywickrama 2010: 2; Berger 2014: 63; Bukta 2014: 113). In terms of transparency, the necessity becomes more evident when the issue is viewed from the students' perspective. As already mentioned in section 2.1., EFL writers often concentrate more on the content, find it difficult to focus on different textual features at the same time and frequently do not realise that continuous work is necessary to reach a higher level (Hyland 2003: 10-11). Thus, the language teachers must provide an understandable and transparent feedback tool to give their students enough opportunities to be able to actually make progress.

7. Conclusion

In conclusion, important observations were made that were connected to teaching and testing in the EFL classroom and partly prove the initial hypothesis of this paper.

First of all, a fair, objective and unbiased rating approach could not be seen in this study but at the same time, it could also not exactly be determined that the rating approach of the participating teachers was unfair, subjective or biased. In fact, it is rather difficult to reconstruct the complex process of grading written compositions, and the relationship between the rater, the scale and the text remains unclear. However, the analysis and discussion of the research showed that some incidents of biased grading or feedback support the hypothesis which claimed that the teachers' grading is influenced by their personal opinions or preferences. Evidence for that could be found in cases in which teachers only gave feedback on parts of the assessment criteria or a biased approach in the rating process could be proven by comparing the teachers' feedback to the comments on the texts.

The initial research question made in the introduction of this paper asked whether teachers who grade with the analytic rating scale and who underwent rater training for writing are still somehow affected in their grading process by influences like personal experience, opinions or beliefs about good writing. This question could partly be answered with a yes. The assumptions made about the insights gained from the teachers' written feedback on the exam texts must be treated with caution as they are not entirely conclusive. The participants' answers to the statements in the questionnaire still display a few indications of biased grading or inaccurate use of the assessment scales. In cases where the questionnaire answers suggested a biased application of the scale and the graded texts supported this theory, a positive answer to the research question is most evident. Although the reasons for the teachers' choices might not always be clear, the few instances in which there is an obvious tendency towards one of the assessment criteria support this outcome. Similar observations were made in cases in which the scale is definitely used wrong or the scale's analytic character is ignored.

As the CLAAS was created to give detailed feedback for students' continuous work and the improvement of their writing skills, applying it correctly is necessary to benefit from the opportunity to use it as a washback tool. Students can gain a complex insight into their strengths and weaknesses and can learn to manage their writing process to become skilled writers. For the raters that means using the scale, understanding the scale and its guidelines and learning how to apply it in an analytical way. Although the teachers in this study all underwent rater training, there are still discrepancies in the use of the CLAAS. The results of the study definitely show that more rater training is necessary for the teachers to become more proficient in the use of the scale and to ensure a more objective, unbiased and fair rating approach which follows the analytic character of the scale.

25660 words

8. References

Bachmann, Lyle F.; Palmer. Adrian S. 1996. *Language testing in practice: designing and developing useful language tests*. Oxford: Oxford University Press.

Baker, Beverly Anne. 2012. “Individual differences in rater decision-making style: an exploratory mixed-methods study”. *Language Assessment Quarterly* 9(3), 225-248.

Barkaoui, Khaled. 2010. “Variability in ESL essay rating processes: the role of the rating scale and rater experience”. *Language Assessment Quarterly* 7(1), 54-74.

Barkaoui, Khaled. 2011. “Think-aloud protocols in research on essay rating: an empirical study of their veridicality and reactivity”. *Language Testing* 28(1), 51-75.

Belludi, Nagesh. 2010. “The halo and horns effect [rating errors]”. http://www.rightattitudes.com/2010/04/30/rating-errors-halo-effect-horns-effect (11 Dec 2018).

BIFIE. 2014a. “Assessment scale B2 and guidelines”. Wien: Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens. https://www.srdp.at/downloads/dl/beurteilungsraster-b2-deutsch-und-englischsprachig-und-guidelines-1/ (16 July 2018).

BIFIE. 2014b. “Beurteilungsraster B1 und Begleittext [Assessment scale B1 and guidelines]”. Wien: Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens. https://www.srdp.at/downloads/dl/beurteilungsraster-b1-und-begleittext/ (16 July 2018).

BIFIE. 2018. “Über das BIFIE [About BIFIE]”. Wien: Bundesinstitut für Bildungsforschung, Innovation & Entwicklung des österreichischen Schulwesens. https://www.bifie.at/ueber-bifie/organisation/ (16 July 2018).

Brown, H. Douglas; Abeywickrama, Priyanvada. 2010. *Language assessment: principles and classroom practices*. (2nd ed.) New York, NY: Pearson Longman.

Bukta, Katalin. 2014. *Rating EFL written performance*. Warsaw: de Gruyter.

BMBWF. n.d. “Durchführung [Procedure]” *Standardisierte Reife-und Diplomprüfung Page*. https://www.srdp.at/durchfuehrung/ (11 Dec. 2018).

Bundesministerium für Bildung. 2013. “Kommentierte Schreibperformanz in Englisch [Commented writing performance in English] ”. *Standardisierte Reife- und Diplomprüfung*. https://www.srdp.at/index.php?eID=dumpFile&t=f&f=1200&token=a959ca18e5457732dc666e5ed3406d908a4b1ee1 (11 Dec. 2018).

BMBWF. 2017. *Übersicht Charakteristika Textsorten lebende Fremdsprachen (SRP/AHS, SRDP/BHS) [Overview of text type features in modern foreign languages]*. https://www.srdp.at/index.php?eID=dumpFile&t=f&f=2645&token=0807a51b2cee53be41d94460e6e3235a8d619941 (15 Dec. 2018).

BMBWF. 2018. *Standardisierte kompetenzorientierte Reifeprüfung an AHS [Standardised competence-oriented school-leaving exam at grammar schools].* https://bildung.bmbwf.gv.at/schulen/unterricht/ba/reifepruefung.html#heading_S_ule_2_Klausurarbeiten_ (13 Dec. 2018).

Cohen, Andrew, D. 1994. *Assessing language ability in the classroom*. Boston, MA: Heinle & Heinle.

Congdon, Peter J.; McQueen, Joy. 2000. "The stability of rater severity in large-scale assessment programs". *Journal of Educational Measurement* 37(2), 163–178.

Council of Europe. 2001. *Common European framework of reference for languages*. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf (11 Dec. 2018)

Cumming, Alister; Kantor, Robert; Powers, Donald E. 2002. "Decision making while rating ESL/EFL writing tasks: a descriptive framework". *Modern Language Journal* 86(1), 67-96.

Dörnyei, Z.; Csizér, K. 2012. "How to design and analyze surveys in second language acquisition research". In Mackey, Alison; Gass, Susan M. (eds.). *Research methods in second language acquisition: a practical guide*. Chichester: Blackwell, 74-94.

Elbow, Peter. 1996. "Writing assessment: do it better, do it less". In White, Edward M.; Lutz, William D.; Kamusikiri, Sandra (eds.). *Assessment of writing: politics, policies, practices*. New York; NY: The Modern Language Association of America, 120-34.

Engelhard, George Jr. 1994. "Examining rater errors in the assessment of written composition with a many-faceted Rasch model". *Journal of Educational Measurement* 31(2), 93-112.

Feeley, Thomas Hugh. 2002. "Comments on halo effects in rating and evaluation research". *Human Communication Research* 28(4), 578-586.

Flower, Linda. 1989. "Cognition, context and theory building". *College Composition and Communication* 40, 282-311.

Flower, Linda., Hayes, John. 1981. "A cognitive process theory of writing". *College Composition and Communication* 32, 365-87.

Goulden, Nancy Rost. 1992. "Theory and vocabulary for communication assessments". *Communication Education* 41(3), 258-269.

Hamp-Lyons, Liz. 1991. "Scoring procedures for ESL contexts". In Hamp-Lyons, Liz (Ed.), *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex, 241–276.

He, Tung-Hsien; Gou, Wen Johnny; Chien, Ya-Chen; Chen, I-Shan Jenny, Chang, Shan-Mao. 2013. "Multi-faceted Rasch measurement and bias patterns in EFL writing performance assessment". *Psychological Reports* 112(2), 469-485.

Hughes, Arthur. 2003. *Testing for language teachers*. (2nd ed.). Cambridge: Cambridge University Press.

Hyland, Ken. 2003. *Second language writing*. Cambridge: Cambridge University Press.

Lance, Charles E.; Lapointe, Julie A.; Stewart, Amy M. 1994. "A test of the context dependency of three causal models of halo rater error". *Journal of Applied Psychology*, 79(3), 332-340.

Lopatina, Kristina. 2016. "Implementierung der Zentralmatura in Österreich: Akzeptanz bei Lehrkräften [Implementation of the new matura in Austria: acceptance among teachers]". MA thesis, University of Vienna.

Lumley, Tom. 2002. "Assessment criteria in a large-scale writing test: what do they really mean to the raters?". *Language Testing* 19(3), 246-276.

Malouff, John M.; Thorsteinsson, Einar B. 2016. "Bias in grading: a meta-analysis of experimental research findings". *Australian Journal of Education* 60(3), 245-256.

McNamara, Tim Francis. 1996 *Measuring second language performance*. New York, NY: Longman.

Murphy, Kevin R., Cleveland, Jeanette N. 1991. *Performance appraisal: an organizational perspective*. Boston, MA: Allyn & Bacon.

Raimes, Ann. 1983. *Techniques in teaching writing*. New York, NY: Oxford University Press.

Saal, Frank E.; Downey, Ronald G.; Lahey, Mary A. 1980. "Rating the ratings: assessing the psychometric quality of rating data". *Psychological Bulletin* 88(2), 413-428.

Taschwer Klaus, Illetschko Peter. 2018. "Mathematik-Matura: Was Experten über die Aufgaben sagen [The school-leaving exam in mathematics: what experts say about the tasks]". *Der Standard Online Edition,* 10 May. https://derstandard.at/2000079528711/Mathematik-Matura-Was-Experten-ueber-die-Aufgaben-sagen (16 July 2018).

Tribble, Christopher. 1996. *Writing*. Oxford: Oxford University Press.

Upshur, John A., & Turner, Carolyn. E. (1995). "Constructing rating scales for second language tests". *ELT Journal*, 49(1), 3-12.

Vaughan, Caroline 1993. "Holistic assessment: what goes on in the rater's mind?" In L. Hamp-Lyons (ed.), *Assessing second language writing in academic contexts*. (2nd edition). Norwood, NJ: Ablex, 111-26.

Weigle, Sara Cushing. 1998. "Using FACETS to model rater training effects". *Language Testing* 15(2), 263-87.

Weigle, Sara Cushing. 2002. *Assessing writing*. Cambridge: Cambridge University Press.

9. Appendices

9.1. Questionnaire sample

Teachers' opinions & personal preferences in connection with BIFIE's assessment scales

The sentences below are all concerned with the grading process of B1 and B2 texts in the context of testing and assessment with the help of the official rating scales (BIFIE) in use. My research deals with the teachers' use of the assessment scales and the relationship between their opinions or preferences about the scales and their use.

First, general and background information is elicited and then the questionnaire asks for more detailed information. The survey will take roughly 10 minutes. All the answers will be treated confidentially, and participants remain anonymous. Tick the box that most closely matches your opinion. Thank you for participating.

General & background information

Name:

Gender:

- ○ Female
- ○ Male

Teaching experience in years:

I have undergone rater training

for writing:

- ○ No
- ○ Yes (Where? How long? ______________________________________)

I have undergone rater training

for speaking:

- ○ No
- ○ Yes (Where? How long? ______________________________________)

As a quick reminder, the abbreviations used in this survey are as follows:

TA= Task Achievement, OL= Organisation & Layout, LSR= Lexical & Structural Range, LSA= Lexical and Structural Accuracy

	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
A very important language aspect of writing is grammatical accuracy.					

I am aware of the guidelines that BIFIE provides for working with the assessment scales.					
I am a rather lenient rater.					
I find it hard to match my impressions of a text with the wording of the scale.					
I find BIFIE’s assessment scales fair.					
Missing markers which show the relationship between the pupil’s ideas are not an exclusive OL problem but should also result in the choice of a lower band in the LSR criterion.					
With every criterion, I start with band 6 and then go up or down the bands, considering the pupil’s writing.					
Grading is a stressful process for me.					
If the word count is not observed, I automatically deduct points in other criteria as well.					
Avoiding repetition of structures is highly important for me.					
I tend to end up in the middle section of the bands.					
If the text is awarded a low band for OL because it is confusing, it cannot receive a good grade overall.					
The use of correct but simple structures must lead to the choice of a lower band in the LSR criterion.					
When a pupil scores low in a criterion, I tend to deduct more points in other criteria as well.					
As long as I can justify my rating choices with the scale, I can freely choose a higher or lower band than others might do.					
The content points (their degree of development) are more important than the other three criteria.					
The wording of the scale is open to interpretation.					
When I was in school, my own texts were graded using analytic rating scales.					
In my opinion, BIFIE’s analytic assessment scales (for level B1/B2) are useful and necessary.					
The task must be fully achieved to get a good grade overall.					
I am aware of the laws and guidelines in the LBVO concerning the grading of written texts.					
Whenever I cannot decide between two bands, I choose the higher one.					
My texts in school were not graded with an analytic rating scale and I found it rather unfair.					
Missing markers which show the relationship between the pupil’s ideas are not an exclusive OL problem but should also result in the choice of a lower band in the LSA criterion.					
When I have an impression of an aspect of the text, I try to find suitable wording in the assessment scale to award a particular band in the relevant criteria.					
A very important language aspect of writing is lexical accuracy.					
A pupil cannot receive very high points in one criterion and very few or zero in another one.					
I would rather correct the texts using my own assessment criteria.					

As long as the task is fulfilled, the writing does not have to be highly accurate in terms of LSA.					
When in doubt, I rather choose the middle section of the bands in each criterion.					
Ignoring paragraph conventions can lead to the choice of a lower band not only in the OL, but also in the TA criterion as one might not be able to follow the thought process.					
Avoiding repetition of vocabulary is highly important for me.					
When a pupil scores high in one criterion, I tend to assign more points to others as well.					
A very important language aspect of writing is lexical range.					
The use of correct but simple structures must lead to the choice of a lower band in the LSA criterion.					
I always read a text more than three times.					
Every mistake is only counted once in each criterion in the correction process.					
A very important language aspect of writing is grammatical range.					
In my grading process, I consider the first descriptors of each band as more important and thus they weigh more in the decision-making process.					

9.2. Questionnaire results

General & background information	Teacher P	Teacher T	Teacher S	Teacher N	Teacher A	Teacher J	Teacher R
Gender	Female	Male	Female	Female	Female	Female	Male
Teaching experience in years	3	8	2	7	27	11	20
I have undergone rater training for writing.	Yes (1 afternoon SCHILF Fortbildung)	Yes (University course, 1 semester; further teacher training courses)	Yes (1 afternoon in St.Pölten Fortbildung)	Yes (@school with Prof.A.Berger 1 afternoon)	Yes (Hollabrunn 2 days, Vienna ½)	Yes (PH NÖ 4 UE + 4 UE)	Yes (various seminars)
I have undergone rater training for speaking.	No	Yes (University course, 1 semester; further teacher training courses)	No	Yes (@school with Prof.A.Berger 1 afternoon)	Yes (Hollabrunn, 1 day)	Yes (PH NÖ 5 UE)	Yes (various seminars)

Questionnaire	Teacher P	Teacher T	Teacher S	Teacher N	Teacher A	Teacher J	Teacher R	Average	Standard deviation
A very important language aspect of writing is grammatical accuracy.	4	4	4	4	3	2	4	3.6	0.7
I am aware of the guidelines that BIFIE provides for working with the assessment scales.	5	5	4	5	5	2	5	4.3	1.1
I am a rather lenient rater.	3	2	2	3	3	1	3	2.4	0.7

I find it hard to match my impressions of a text with the wording of the scale.	3	2	5	2	4	4	2	3.1	1.1
I find BIFIE’s assessment scales fair.	4	4	2	4	4	4	4	3.7	0.7
Missing markers which show the relationship between the pupil’s ideas are not an exclusive OL problem but should also result in the choice of a lower band in the LSR criterion.	3	2	3	2	2	5	3	2.9	0.9
With every criterion, I start with band 6 and then go up or down the bands, considering the pupil’s writing.	4	5	5	5	4	5	3	4.4	0.7
Grading is a stressful process for me.	4	4	5	2	4	4	3	3.7	0.9
If the word count is not observed, I automatically deduct points in other criteria as well.	3	2	4	2	1	2	2	2.3	0.9
Avoiding repetition of structures is highly important for me.	4	4	3	4	4	4	3	3.7	0.5
I tend to end up in the middle section of the bands.	2	3	2	3	2	4	1	2.4	0.9
If the text is awarded a low band for OL because it is confusing, it cannot receive a good grade overall.	2	4	3	4	4	3	3	3.3	0.7
The use of correct but simple structures must lead to the choice of a lower band in the LSR criterion.	3	4	4	2	3	5	4	3.6	0.9
When a pupil scores low in a criterion, I tend to deduct more points in other criteria as well.	1	2	2	2	3	2	1	1.9	0.6

As long as I can justify my rating choices with the scale, I can freely choose a higher or lower band than others might do.	3	4	3	3	3	4	3	3.3	0.5
The content points (their degree of development) are more important than the other three criteria.	2	3	3	2	2	3	2	2.4	0.5
The wording of the scale is open to interpretation.	4	4	5	5	4	4	5	4.4	0.5
When I was in school, my own texts were graded using analytic rating scales.	1	2	1	1	1	1	1	1.1	0.3
In my opinion, BIFIE's analytic assessment scales (for level B1/B2) are useful and necessary.	4	5	4	4	4	4	4	4.1	0.3
The task must be fully achieved to get a good grade overall.	3	4	5	4	4	3	4	3.9	0.6
I am aware of the laws and guidelines in the LBVO concerning the grading of written texts.	5	5	3	5	5	4	5	4.6	0.7
Whenever I cannot decide between two bands, I choose the higher one.	2	2	3	4	3	2	3	2.7	0.7
My texts in school were not graded with an analytic rating scale and I found it rather unfair.	4	3	4	4	4	2	1	3.1	1.1
Missing markers which show the relationship between the pupil's ideas are not an exclusive OL problem but should also result in the choice of a lower band in the LSA criterion.	4	2	4	2	2	1	1	2.3	1.2

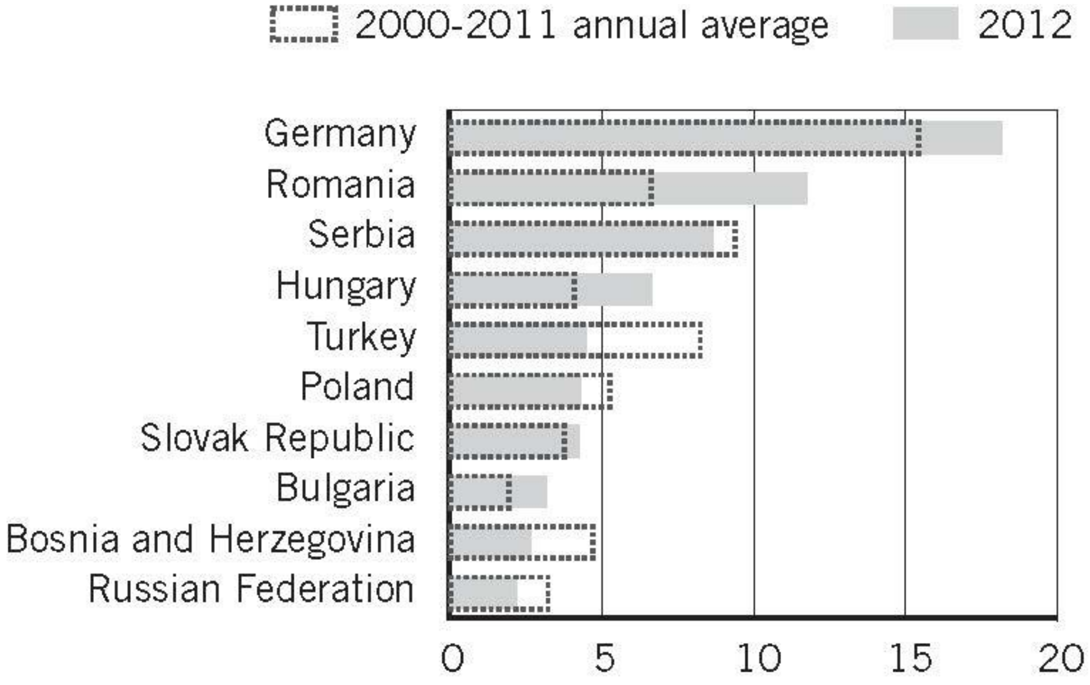
When I have an impression of an aspect of the text, I try to find suitable wording in the assessment scale to award a particular band in the relevant criteria.	3	4	4	3	4	5	3	3.7	0.7
A very important language aspect of writing is lexical accuracy.	4	4	4	4	4	2	4	3.7	0.7
A pupil cannot receive very high points in one criterion and very few or zero in another one.	2	4	4	5	2	1	4	3.1	1.4
I would rather correct the texts using my own assessment criteria.	2	2	3	3	2	1	1	2.2	0.7
As long as the task is fulfilled, the writing does not have to be highly accurate in terms of LSA.	2	2	2	2	2	2	4	2.3	0.7
When in doubt, I rather choose the middle section of the bands in each criterion.	2	3	3	2	3	3	3	2.7	0.5
Ignoring paragraph conventions can lead to the choice of a lower band not only in the OL, but also in the TA criterion as one might not be able to follow the thought process.	4	4	4	4	3	2	4	3.6	0.7
Avoiding repetition of vocabulary is highly important for me.	4	4	4	4	5	5	4	4.3	0.5
When a pupil scores high in one criterion, I tend to assign more points to others as well.	2	2	2	2	2	2	4	2.3	0.7
A very important language aspect of writing is lexical range.	4	4	5	4	4	4	5	4.3	0.5

The use of correct but simple structures must lead to the choice of a lower band in the LSA criterion.	3	4	2	2	5	1	4	3	1.3
I always read a text more than three times.	5	1	3	1	4	2	1	2.4	1.5
Every mistake is only counted once in each criterion in the correction process.	5	3	4	4	3	4	2	3.6	0.9
A very important language aspect of writing is grammatical range.	4	4	4	4	4	3	3	3.7	0.5
In my grading process, I consider the first descriptors of each band as more important and thus they weigh more in the decision-making process.	2	2	4	2	3	2	1	2.3	0.9

9.3. Prompts and graded texts

Teacher T's prompt for the writing task in an exam of the 8th grade:

Your class is doing a project on immigration into EU countries in recent years. You have decided to write about the situation in Austria and have found the following data about the top 10 nationalities of immigrants as a percentage of total immigrant arrivals in Austria, 2000–2012.

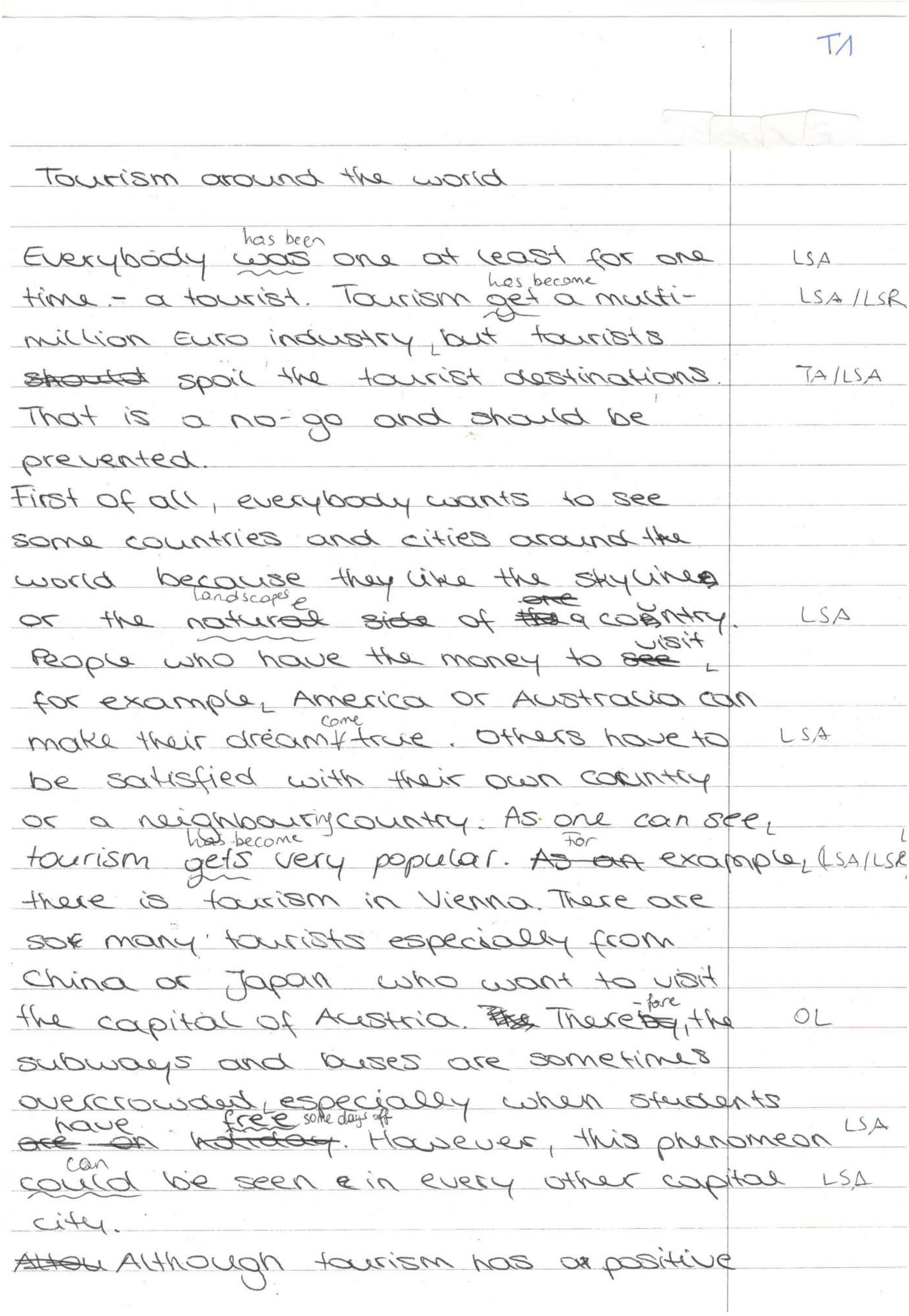


In your report to the programme coordinator, Mrs. Diane Olsen, you should:

- analyse the chart
- discuss any significant trends
- comment on the reasons for immigration into Austria.

Write around 250 <u>OR</u> 400 words.
Divide your report into sections and give them headings.

Teacher T's graded exam texts, T1 and T2:



T1

Tourism around the world

Everybody has been was one at least for one LSA
time - a tourist. Tourism has become get a multi- LSA/LSR
million Euro industry, but tourists
~~should~~ spoil the tourist destinations. TA/LSA
That is a no-go and should be
prevented.
First of all, everybody wants to see
some countries and cities around the
world because they like the skyline~~s~~
or the landscapes ~~natureal~~ ~~side~~ of ~~the~~ ~~one~~ a country. LSA
People who have the money to ~~see~~ visit,
for example, America or Australia can
make their dream~~s~~ come true. Others have to LSA
be satisfied with their own country
or a neighbouringcountry. As one can see,
tourism has become gets very popular. ~~As an~~ For example, LSA/LSR
there is tourism in Vienna. There are
so many tourists especially from
China or Japan who want to visit
the capital of Austria. ~~The~~ There-fore~~by~~, the OL
subways and buses are sometimes
overcrowded, especially when students
~~are on~~ have free some days off ~~holiday~~. However, this phenomeon LSA
can could be seen ~~e~~ in every other capital LSA
city.
~~Allou~~ Although tourism has ~~a~~ positive

T1

LSA effect on the economy, there are some parts discussed.
LSA, LSA that ~~they~~ have to be ~~discussed~~ On the
environment tourism has ~~as~~ negative
LSA effect because every tourist needs a ~~son~~
LSA/LSR vehicle to reach to get on holiday his destination.
No matter if this is a car, train or
LSA, LSA, LSA LSR plane. Everyone using these means of transport ~~of this~~ causes air
pollution and that is informal! bad for our
environment. However, not only the
global environment suffers from tourism
LSA but also the tourist destinations themselves (~~itself~~).
Unfortunately, there are some tourists
who draw some figures or quotes
on ~~a~~ walls of ~~som~~ famous places.
This makes the place not as special
as it was before.
Also ~~co~~ society can suffer from tourism
LSA because segregation ~~For example tourism bri~~ can be cause. There
LSA are some races that are thought to be hold for
dangerous because of ~~the~~ wars in
LSA their countries. ~~Someone~~ One does not have
to forget that many human beings
~~do not~~ are not guilty because it is not
their fault.
no direct questions in essays! → TA So what can be done to prevent
the ~~some~~ negative effects of tourism.
First, everbody has to be open and ~~to~~
LSA accept other ethnic groups. We are
LSA all human and we want all of us to be

treated well. If we accept each other LSA
there would be neither war nor segregation. LSA
Moreover, ~~Also~~ the vandalism in ~~damage on~~ several places and OL, LSR
buildings has to be stopped. Nobody LSA
should draw or cut something on or into
walls. If everybody did ~~would do~~ that, many LSA
cities would be ~~were~~ colorful. However, we LSA
do not want colorful cities but (LSA)
beautiful and historical places. Maybe | that easy? → TA
a sign would prevent vandalism. In
my opinion, it would be useful to
make nearer cities attractive so that
not every person travels with the plane (LSA)
to another country but with the car
to ~~in~~ the next city. LSA
To sum ~~of these things~~ up, one can OL/LSR
say that tourism has both advantages
and disadvanteges. We do not have to
forget that we are guests in another
city and that we are not allowed to
spoil everything.

TA: 6/10
OL: 6/10
LSR: 4/10
LSA: 2/10
18/40

~420 words

T2

1st Test

15. December 201

250 words

IMMIGRATION INTO AUSTRIA

To: Mrs. Diane Olsen, programme coordinator
From: A[redacted] D[redacted]
Date: 15th of December, 2017
Subject: Immigration into Austria

Introduction

Because of our class project I have decided to write this report about the immigration into Austria. To do so, I chose to work with ~~this~~ a chart about immigration showing data from 2000 till 2012.

LSA/TA

Findings

The bar graph shows the 10 most common countries (that) people come from before moving to Austria. Each country has two graphs; one shows the average immigration percentage between 2000 and 2011 and the other one the percentage of immigrants in 2012. people believe tha~~t~~ all the Although most ~~people~~ immigrants come from Africa or East Asia, the top 10 nationalities to immigrate into Austria are all European.

OL

LSA

T2 A. F

15 percent, the ~~biggest amount~~ highest number, of LSA
immigrants actually ~~come from~~ are our
(~~fellow~~) neighbours, the Germans. LSA
Also important to mention is that
only four countries' immigration to Austria is TA
increasing. ~~The bigg~~

Significant Trends

The biggest growth in immigrants ~~[illegible]~~
~~[illegible]~~ is from Romania. The average from
eleven years ~~is~~ differs from the results
from one year, 2012, in as much as
five percent of total immigration into
Austria.
~~German immigration~~ Only half the
amount of people immigrated from Turkey
in 2012, not even reaching five percent.
The annual average 200 between 2000 and
2011 ~~[illegible]~~ where about eight percent. *

Reasons and conclusion OL/TA

Austria is a very popular country, because
it's economy is very stable and the living LSA
standards are also very high. The reason
why almost twenty percent of all immigrants
are from Germany could be that our
university tuition fees are much cheaper than LSA
(they are) in Germany.

T2 A.I

*The smallest (~~amount of~~) change between
for the figures of
the two graphs happened in the
Slovak Republic. The percentage only
dropped about half a percent, making
quite
Slovakian immigration pretty steady.

LSA
TA/LSA
LSA, LSR

262 words

TA: 9/9
OL: 9/9
LSR: 9/9
LSA: 9/9
36/40

Teacher S’s prompt for the writing task in an exam of the 5th grade:

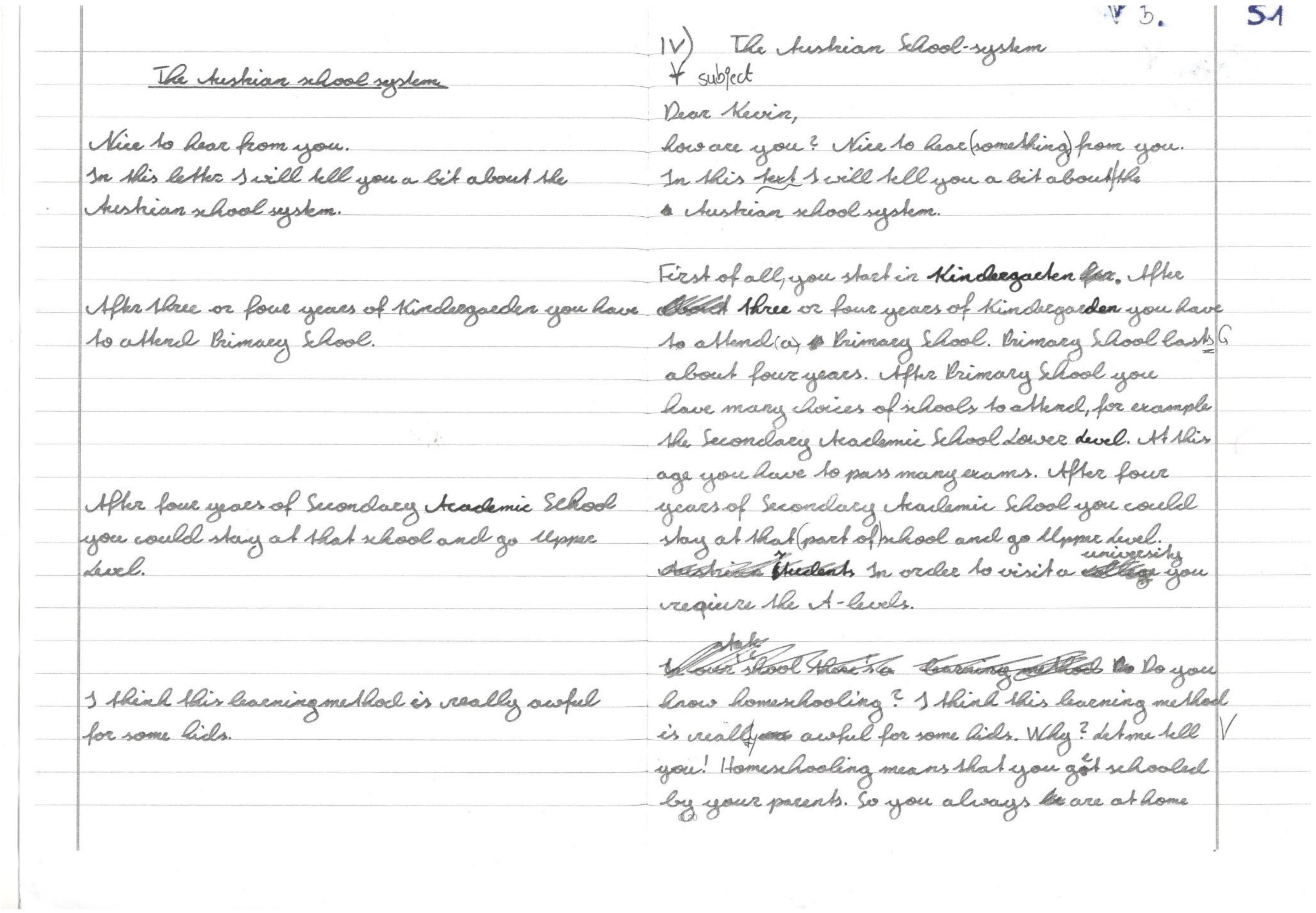
IV) WRITING (40%)

Write an informal email to a friend, where you

- Describe the Austrian school system
- Explain an unusual learning method
- Discuss what your ideal school would look like.

Write around 200 words (+/- 10%)

Teacher S's graded exam texts, S1 and S2:



The Austrian school system

Nice to hear from you.
In this letter I will tell you a bit about the
Austrian school system.

After three or four years of Kindergarden you have
to attend Primary School.

After four years of Secondary Academic School
you could stay at that school and go Upper
level.

I think this learning method is really awful
for some kids.

V 3.

S1

IV) The Austrian School-system
subject

Dear Kevin,
how are you? Nice to hear (something) from you.
In this ~~text~~ I will tell you a bit about the
Austrian school system.

First of all, you start in Kindergarten ~~for~~. After
~~About~~ three or four years of Kindergarden you have
to attend (a) Primary School. Primary School lasts G
about four years. After Primary School you
have many choices of schools to attend, for example
the Secondary Academic School Lower level. At this
age you have to pass many exams. After four
years of Secondary Academic School you could
stay at that (part of) school and go Upper level.
~~Austrian~~ Students In order to visit a ~~college~~ university you
require the A-levels.

~~In our school there's a~~ state ~~learning method~~ Do you
know homeschooling? I think this learning method
is really ~~are~~ awful for some kids. Why? Let me tell V
you! Homeschooling means that you get schooled
by your parents. So you always are at home

S1

because you don't have to go to school. I think that's really awful because you don't meet new friends and you always have to learn alone. What's your opinion about that?

~~A ideal school would be For me a ideal school would be~~ An ideal school is that you get have nice G

teachers and you learn these ~~[illegible]~~ ≠ which you G,G

want to.

So that's all. Have a nice weekend!

Yours, Daniel

6/6/5/6

An ideal school is that you have nice teachers and you learn these subjects which you want to.

Q 20

52

Correction

subject: school system
I am writing to discuss the Austrian school system.

After that you can choose between middle school and high school, that are secondary schools.
After those four years you can go to vocational middle school and graduate or do an apprenticeship.

I think it is a little bit scary and I cannot imagine that the robot could work as well as a real teacher.
The real teacher can better help you with your individual problems and a robot cannot.

IV) ≠ subject!!

Dear Lisa,

I am writing to discuss ~~about~~ the G
Austrian school system.
At the age of six to nine, you start with
primary school. After that you can
choose between "Hauptschule" and V
"Gymnasium", that are ~~the~~ secondary V
schools. After those four years
you can go to the "berufsbildende V
mittlere Schule" and graduate, or do an apprenticeship +
and start to work.

I would also like to know what you
think about a virtual teacher.
I think it is a little bit scary
and I cannot imagine that it could ?
work as ~~good~~ well as a real teacher. G
The real teacher can help you
better with your individual problems
and a robot~~ex~~ cannot. And if you do not V
understand something, the real teacher

Q 20

nd if you do not understand something, .real teacher can help you better.

I think it is the best because you can ften choose which way, you want to o and for example in the American chool system you cannot choose as ften.
Another thing is that in Austria you ave to go to school only for nine years and then you can go to work.

52

can help you better. Could you please let me know if there could be a change in this topic?

|| too formal (→ register)

? what kind of change? (logic!)

My ideal school would look like the Austrian school system. I think it is the best because you can often choose which way you want to go and for example in the American school system you cannot choose as often (as ~~by~~ us). Another thing is that in Austria you have to ~~have~~ go to school only ~~for~~ nine years and then you can go to work. In America, you have to go to school for a longer time.

G
G

I look forward to hearing from you

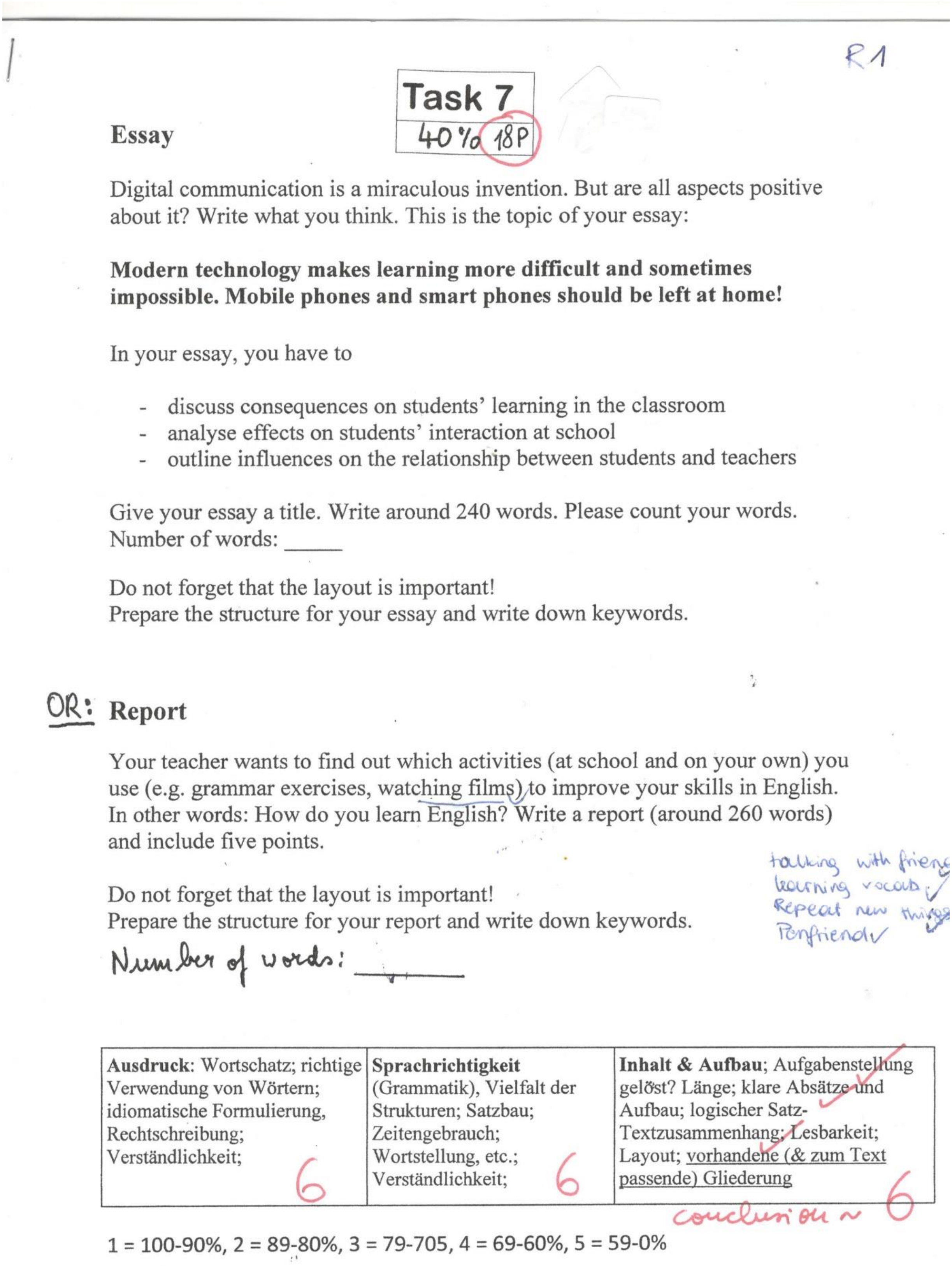
Yours faithfully

||→ register!

Julie Vòlld

4/5/4/6

Teacher R's prompt for the writing task in an exam of the 6th grade & the graded exam texts, R1 and R2:



|

R1

Task 7

40% 18P

Essay

Digital communication is a miraculous invention. But are all aspects positive about it? Write what you think. This is the topic of your essay:

Modern technology makes learning more difficult and sometimes impossible. Mobile phones and smart phones should be left at home!

In your essay, you have to

- discuss consequences on students' learning in the classroom
- analyse effects on students' interaction at school
- outline influences on the relationship between students and teachers

Give your essay a title. Write around 240 words. Please count your words.
Number of words: _____

Do not forget that the layout is important!
Prepare the structure for your essay and write down keywords.

OR: **Report**

Your teacher wants to find out which activities (at school and on your own) you use (e.g. grammar exercises, watching films) to improve your skills in English. In other words: How do you learn English? Write a report (around 260 words) and include five points.

talking with friends
learning vocab ✓
Repeat new things ✓
Penfriend ✓

Do not forget that the layout is important!
Prepare the structure for your report and write down keywords.

Number of words: ______

Ausdruck: Wortschatz; richtige Verwendung von Wörtern; idiomatische Formulierung, Rechtschreibung; Verständlichkeit; 6	**Sprachrichtigkeit** (Grammatik), Vielfalt der Strukturen; Satzbau; Zeitengebrauch; Wortstellung, etc.; Verständlichkeit; 6	**Inhalt & Aufbau**; Aufgabenstellung gelöst? Länge; klare Absätze und Aufbau; logischer Satz-Textzusammenhang; Lesbarkeit; Layout; vorhandene (& zum Text passende) Gliederung 6

conclusion ~

1 = 100-90%, 2 = 89-80%, 3 = 79-705, 4 = 69-60%, 5 = 59-0%

7) Report

R

a foreign language difficult
Learning (English) can be quite hard sometimes (especially in my cas when it's not your mother tongue) But gradually (Through years) I've discovered some ways to improve my English and now to ~~get better~~ ~~in~~ gain some "learning skills".

Penfriends

Maybe the best way to learn a foreign language is to talk or to write a lot in it. I've got my penfriend Dan, who ~~absol~~ absolutely helps me learning new words and phrases. ~~the~~ I exchange letters write him (use to) (write) letters from Austria to New York ~~sir~~ for him) in English of course. I'm always very happy when he sends ~~me~~ a letter back with new informations about how his life has recently changed etc. In his letters he also often gives me tips feedback how to improve my English and I also learn new vocabulary from his writings. I'm very glad that my teacher has started ~~with~~ this project; ~~It's very useful for me.~~ I find it very useful.

I'm glad to hear that!

Talking English with friends

During our time in Dublin, we've got to know a new culture and new people. ~~They~~ So my friends and I decided to talk only in English for a few days. At the beginning it was a little bit challenging, but after a few hours? minutes we ~~were so reggni~~ recognized that talking just in English is quite helpful for our ~~pronunciat~~ ~~pronunciation~~ pronounciation and vocabulary.

You really did?

Repeat ~~ers~~ new words

I think the most important point to become fluent in a ~~new~~ language is to repeat new words and phrases all the time. For example, when I was in Dublin, I've heard ~~no~~ many

R1

strange words, I've never heard before, most of the time from my host family. As I told them I'm not sure about the meaning of a word, they explained it to me and so I've started to use ~~this~~ these words. Did you write them down, too? |

Watching films

On Netflix and ~~In~~ in general I use to watch films and series only in English. It helps me a lot to improve my skills. Sometimes the actors are talking in informal English but with the help of my English teacher, I know which ones to use and which not. It's a really modern and new way to ~~a~~ improve one's (your) ~~s~~ skills.

Learning vocabulary

As I've already mentioned, repeating and learning vocabulary is very important. I think it's the key to ~~suceed~~ succeed in English. Without (the necessary) vocabulary you wouldn't be able to ~~a~~ even understand a ~~A~~ foreign language. I personally use this option of learning English the most.

this report shows...

I hope [now you understand the certain ways I improve my English.

(1-2 more phrases

at the beginning the report is very neutral

There are many other methods and activities but these are the most important ones for me....

It was interesting for me to read this report.

R2

Task 7

40 % 18P

15

Essay

Digital communication is a miraculous invention. But are all aspects positive about it? Write what you think. This is the topic of your essay:

Modern technology makes learning more difficult and sometimes impossible. Mobile phones and smart phones should be left at home!

In your essay, you have to

- discuss consequences on students' learning in the classroom
- analyse effects on students' interaction at school
- outline influences on the relationship between students and teachers

Give your essay a title. Write around 240 words. Please count your words.
Number of words: _____

Do not forget that the layout is important!
Prepare the structure for your essay and write down keywords.

OR: Report

Your teacher wants to find out which activities (at school and on your own) you use (e.g. grammar exercises, watching films) to improve your skills in English. In other words: How do you learn English? Write a report (around 260 words) and include five points.

Do not forget that the layout is important!
Prepare the structure for your report and write down keywords.

Number of words: 324

TV series
helping my brother
Speaking English
reading English newsp
playing video games

Ausdruck: Wortschatz; richtige Verwendung von Wörtern; idiomatische Formulierung, Rechtschreibung; Verständlichkeit;
5(+)
	Sprachrichtigkeit (Grammatik), Vielfalt der Strukturen; Satzbau; Zeitengebrauch; Wortstellung, etc.; Verständlichkeit;
	5
		Inhalt & Aufbau; Aufgabenstellung gelöst? Länge; klare Absätze und Aufbau; logischer Satz-Textzusammenhang; Lesbarkeit; Layout; vorhandene (& zum Text passende) Gliederung ✓ ✓
		5

1 = 100-90%, 2 = 89-80%, 3 = 79-705, 4 = 69-60%, 5 = 59-0%

It was interesting to read your report.

R2

Report

inform my teacher about

I was asked to share my personal top five methods for improving my English:

TV-series

Basically every TV-series I watch nowadays, I try to watch in English. This is (easily possible) really easy due to the many great streaming services like Netflix or Sky. I believe (seeing) watching my favourite actors speak English motivates me ① to learn ing the language ~~even~~ 2 ~~more~~. In action-series where there is a lot going on in one moment you can try to practise understanding people in a noisy envoirement.

you = your teacher

advice

~

Studying with my brother

My brother is only one year younger than me, so when we study together I can help him improve his English and revise the
3 stuff leave I already learned. That way I don't forget important grammar rules that I may have forgotten otherwise.

+

R2

1 ×continue

2 learn more~

-> learn more about sth.

improve one's skills...

3 stuff = derogatory

RZ

Speaking English

Fortunately it's very easy ① nowadays to get to know foreign people. Services like Skype let us talk to them all over the world. I have got~~ten~~ to know some American and English people with whom which I now talk regularily. Over a longer period of time this helps me (to) improve my pronounciation as well as my (vocabulary) knowledge. about v.

+

or simply: my vocabulary

Reading English articles

I like to inform myself about politics and economy by reading English articles. This ~~for me~~ is an excellent way for me to learn more advanced °2 words. Always being informed about global desasters is also a hobby °3 of mine. That way I can connect things I enjoy ×4, with the English language, which motivates me even more to improve (myself). my skills.

?

Video games

All video games I play ~~I have~~ are set to English. Altough I don't play very ~~as~~ often anymore, due to the lack of free time, I think this really helped my English a few years prior to now. 5° In video games you need to know what all the game options mean, which meant it was essential for me to learn the meaning of these words. ×6

✓

I hope I was able to properly (share my thoughts). convey my habits

–1 to 2 more sentences?

RZ

1 people from other countries
native speakers

2° diffcult a level can be advance
words can't

3° preoccupation ?

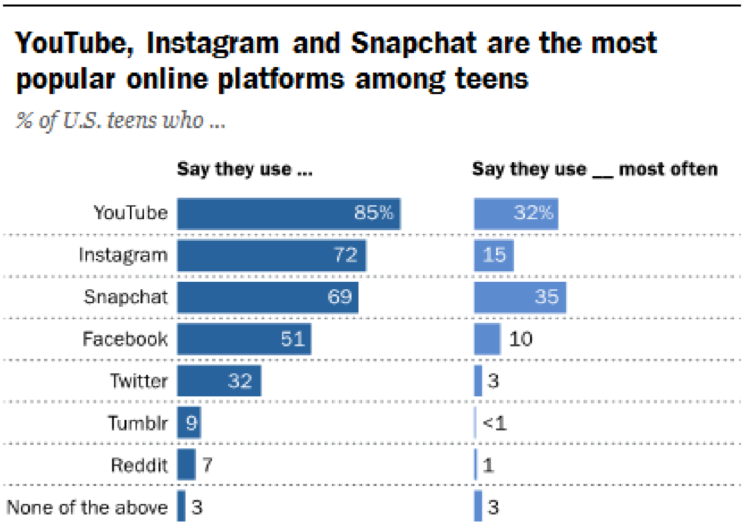
4 × you enjoy reading about
disasters?

5° almost too sticked for a report

6 × slightly long-winded for a very
simple idea

Teacher P’s prompt for the writing task in an exam of the 6th grade:

IV) WRITING (25%)



Note: Figures in first column add to more than 100% because multiple responses were allowed. Questions about most-used site was asked only of respondents who use multiple sites; results have been recalculated to include those who use only one site. Respondents who did not give an answer are not shown.

Source: Survey conducted March 7 –

http://www.pewinternet.org/2018/05/31/teens-social-media-technology-2018/pi_2018-05-31_teenstech_0-01/

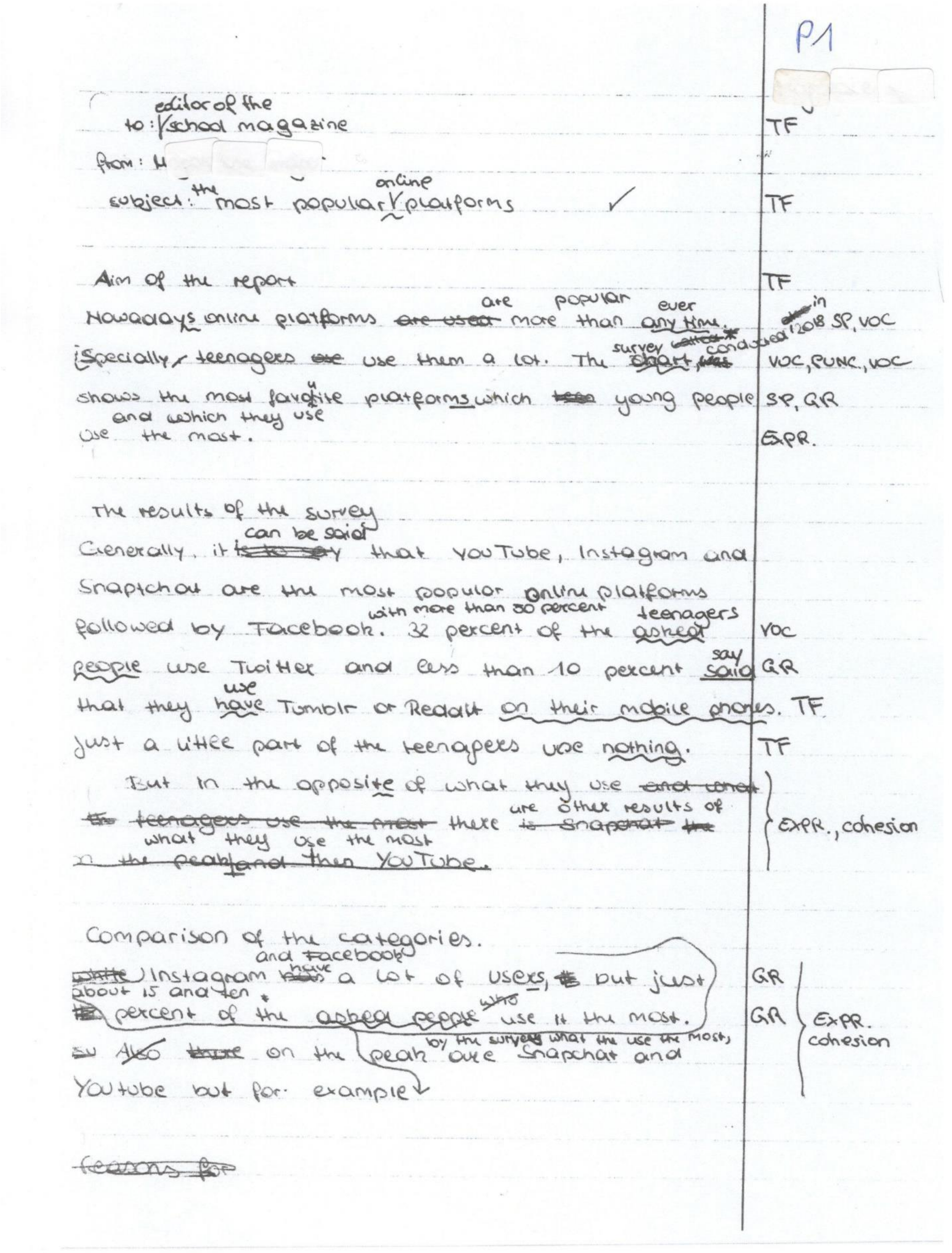
You spend a semester abroad and attend a school in the USA. The last survey on cyberbullying was a real success. So another survey was conducted and you were asked to write a report for the school magazine.

In your report you should:

- Present the results of the survey
- Compare the categories
- Suggest reason for the differences in the two categories

Write about 200 words.

Teacher P’s graded exam texts, P1 and P2:



P1

[redacted]

editor of the
to: school magazine TF

from: H [redacted]

the online
subject: most popular platforms ✓ TF

Aim of the report TF

are popular ever
Nowadays online platforms ~~are used~~ more than any time. ~~[illegible]~~ in 2018 SP, VOC

survey ~~[illegible]~~ * conducted
Specially, teenagers ~~use~~ use them a lot. The ~~chart~~, ~~was~~ VOC, PUNC., VOC

u
shows the most favofite platforms which ~~teen~~ young people SP, GR

and which they use
use the most. EXPR.

The results of the survey

can be said
Generally, it ~~is to say~~ that YouTube, Instagram and

Snaptchat are the most populor online platforms

with more than 30 percent teenagers
followed by Facebook. 32 percent of the asked VOC

say
people use Twitter and less than 10 percent said GR

use
that they have Tumblr or Reddit on their mobile phones. TF

Just a little part of the teenagers use nothing. TF

But in the opposite of what they use ~~and what~~

are other results of
~~the teenagers use the most~~ there ~~is Snapchat the~~ EXPR., cohesion

what they use the most
~~on the peak and then YouTube.~~

Comparison of the categories.

and Facebook have
~~While~~ Instagram ~~has~~ a lot of users, ~~the~~ but just GR

about 15 and ten * who
~~the~~ percent of the asked people use it the most. GR EXPR. cohesion

by the surveys what the use the most,
~~so~~ ~~Also~~ ~~there~~ on the peak are Snapchat and

Youtube but for example

~~Reasons for~~

P1

Reasons for the differences in the two categories

+ ~~Ge~~ Generally ~~you can~~ it ~~can be~~ said that maybe ~~there a platform is used more~~ likely but not u:e ~~the most because~~ people ~~too~~ can give price ~~what~~ more than one platform they use but just one which they use the most. Some teenagers have a lot of apps and also online platforms on their mobile phones and so ~~they~~ some online platforms have a lotf of votes.

EXPR.
ohesion
TF

VOC

GR

GR, GR

236 words

4/5/4/4

~~*'Teens, Social Media & Technology 2018' one which~~

P2

To: editor headmaster of the school magazine TF
From: J
TF

-Aim of this report-
Teenagers People ~~Teenagers~~ were asked about which what online ~~social~~ TF, GR
platforms ~~media~~ ~~websides~~ they use and which ones if they GR
use ~~it~~ most often. There were eight
different platfforms between they could TF
choose, incluoling the opportunity „None of
the above" [Absatz] VOC

- Results of the survey-
The ~~[illegible]~~ survey shows that 85%
uses the ~~ap~~ platfform YouTube and TF EXPR.
out of these, one quarter say said they GR
use it most often. The second largest
community ~~community~~ ~~[illegible]~~ uses Instagram, but only
half as many Instagram users, use
Instagram most often. Snapchat ~~has~~ is
used by more than half of the asked ~~Teenagers~~ People
and ~~[illegible]~~ is one of the teenagers' favourite GR, SP TF, EXPR cohesion
platfforms. "About half of the respondents asked
use facebook and out of them, 10%
~~[illegible]~~ use it ~~the~~ most often. The platfform
twitter has around 30% users.
Less than 4% say said they use Twitter,
Tumblur or Reddi-t ~~[illegible]~~ most often. [Absatz]

- There are two categories - TF
Categorie one shows ~~[illegible]~~ the percentage how many SP TF EXPR.
of teenagers out of the asked people) really using different use the

P2

online platfforms and in the second one | you can see ~~[illegible]~~ ~~one~~

+ ~~[illegible]~~ the percentage of how many

GR people ~~are~~ use ~~using~~ the platfform most often. 162

~~Th~~

TF -The differences between the categories-

~~[illegible]~~ 4. There is a difference between

GR the two graphs ts because not ~~as many~~

~~people~~ everybody uses one social media

TF GR platfform the most. 183 ~~[illegible]~~

The fact that not everybody knows

GR which ~~wich~~ online platfforms they use the

most, causes ~~shows~~ the big difference between

the two graphs. 214

226

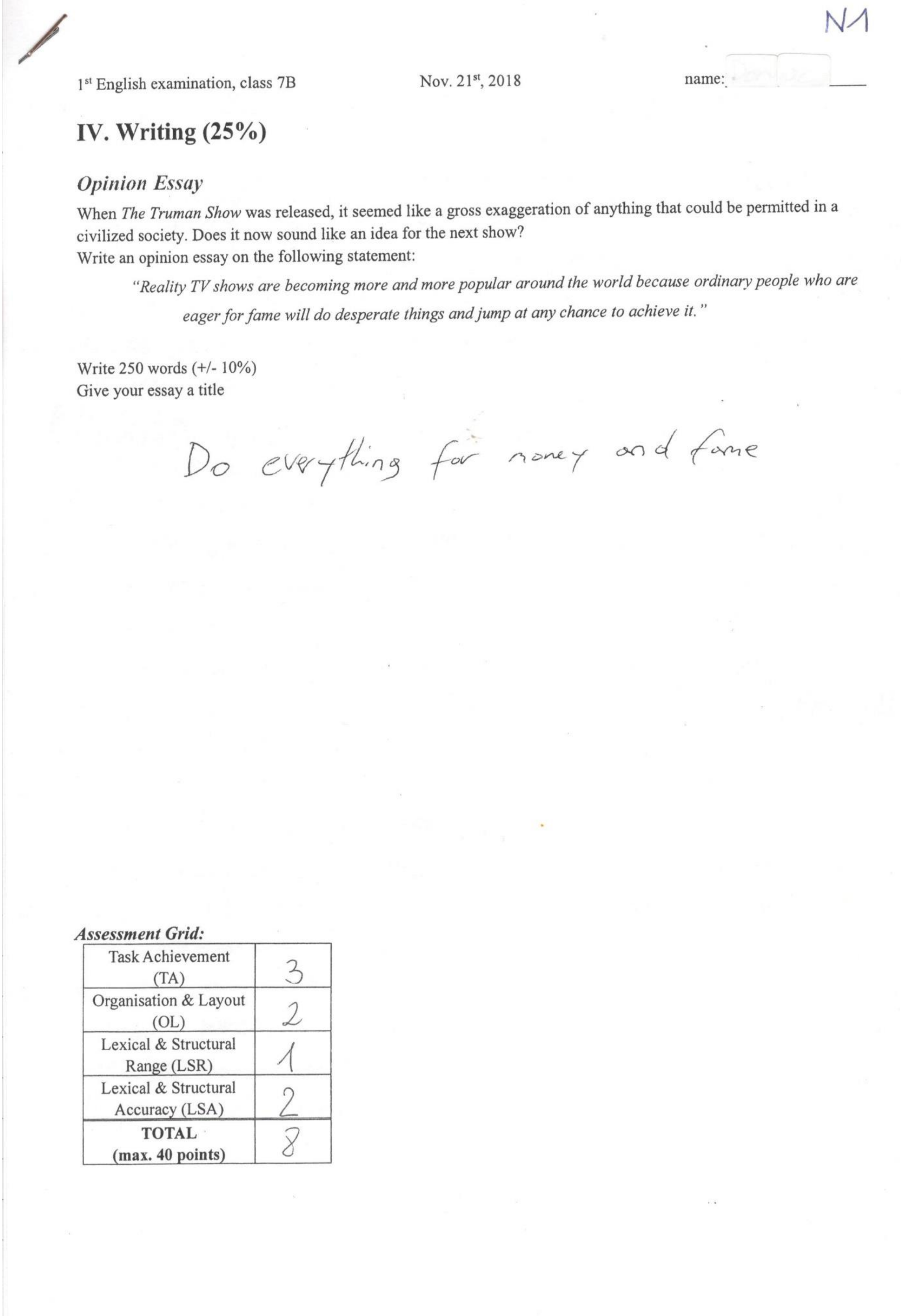
~214

3/4/4/4

[redacted]

99

Teacher N’s prompt for the writing task in an exam of the 7th grade & the graded exam texts, N1 and N2:



N1

1st English examination, class 7B   Nov. 21st, 2018   name: [redacted]

IV. Writing (25%)

Opinion Essay

When *The Truman Show* was released, it seemed like a gross exaggeration of anything that could be permitted in a civilized society. Does it now sound like an idea for the next show?
Write an opinion essay on the following statement:

> *“Reality TV shows are becoming more and more popular around the world because ordinary people who are eager for fame will do desperate things and jump at any chance to achieve it.”*

Write 250 words (+/- 10%)
Give your essay a title

Do everything for money and fame

Assessment Grid:

Task Achievement (TA)	3
Organisation & Layout (OL)	2
Lexical & Structural Range (LSR)	1
Lexical & Structural Accuracy (LSA)	2
TOTAL (max. 40 points)	8

Do everything for money and fame	✓
Is it good for (the) society(,) that every	Expr., P
person s	
famous people get naked (or something like	WW, WF!, Expr
that,) on (the) TV? Now you can see, what	Expr., WF, Exp
the theme of this essay is and maybe you	
for	
can see which side I'm standing (for)	WW, WF
To begin with, the question arises if everybody has	
At the beginning, I want (that) everybody think,	Expr!, WF!
watched a TV show---	
if he or she had watching a TV show like	T!
Adam searching for Eve or "Singlecamp". I mean	P!, Sp., P!
this	
that's very upset to me, at ~~that~~ shows you	WF, WW, P,
can see famous people which needs (a) extra money	WF!, WW
to finance their wonderful life. In the "Singlecamp"	P!
for	
as example, (we see some) people (who) eats or	WW
an animal's testicles	
drinks balls of an animal. That's not helpful	WF!, Expr., WF
for our lifes.	
Secondly, it is impossible for me to	
understand the group of society which	Expr.
(are going to) watch shows like this, every year	T, WO
again. Now my question	Expr.
on this group: "why you are watching this	WW, P, P, T, WO
shows?" Is it interesting or funny to	WW
see famous people do any desperate things	WF!
to earn some money?.	

749

At the end, my question to the producers
of this show is, if they think about the
new TV show twice?. Every child can turn on
Sp., WW the Television and ~~would see~~ might may see,
Expr., WF! some naked people (who trying wants) to find their
P!, WW "big love". Is this in your mind to producers?
NO However, I think also that the producers
WF needs money and aren't better than the (famous) people,
WW! who which are a part of the show.

P, WF, WW Please community, don't watch this obviously
P money making shows. All in all, this message
P, Expr. is very important to parents: (show your kids programs that teach them watch something where your kids can learn) something for their life.

7

N2

1st English examination, class 7B Nov. 21st, 2018 name:________________

IV. Writing (25%)

Opinion Essay

When *The Truman Show* was released, it seemed like a gross exaggeration of anything that could be permitted in a civilized society. Does it now sound like an idea for the next show?
Write an opinion essay on the following statement:

> *"Reality TV shows are becoming more and more popular around the world because ordinary people who are eager for fame will do desperate things and jump at any chance to achieve it."*

Write 250 words (+/- 10%)
Give your essay a title

Assessment Grid:

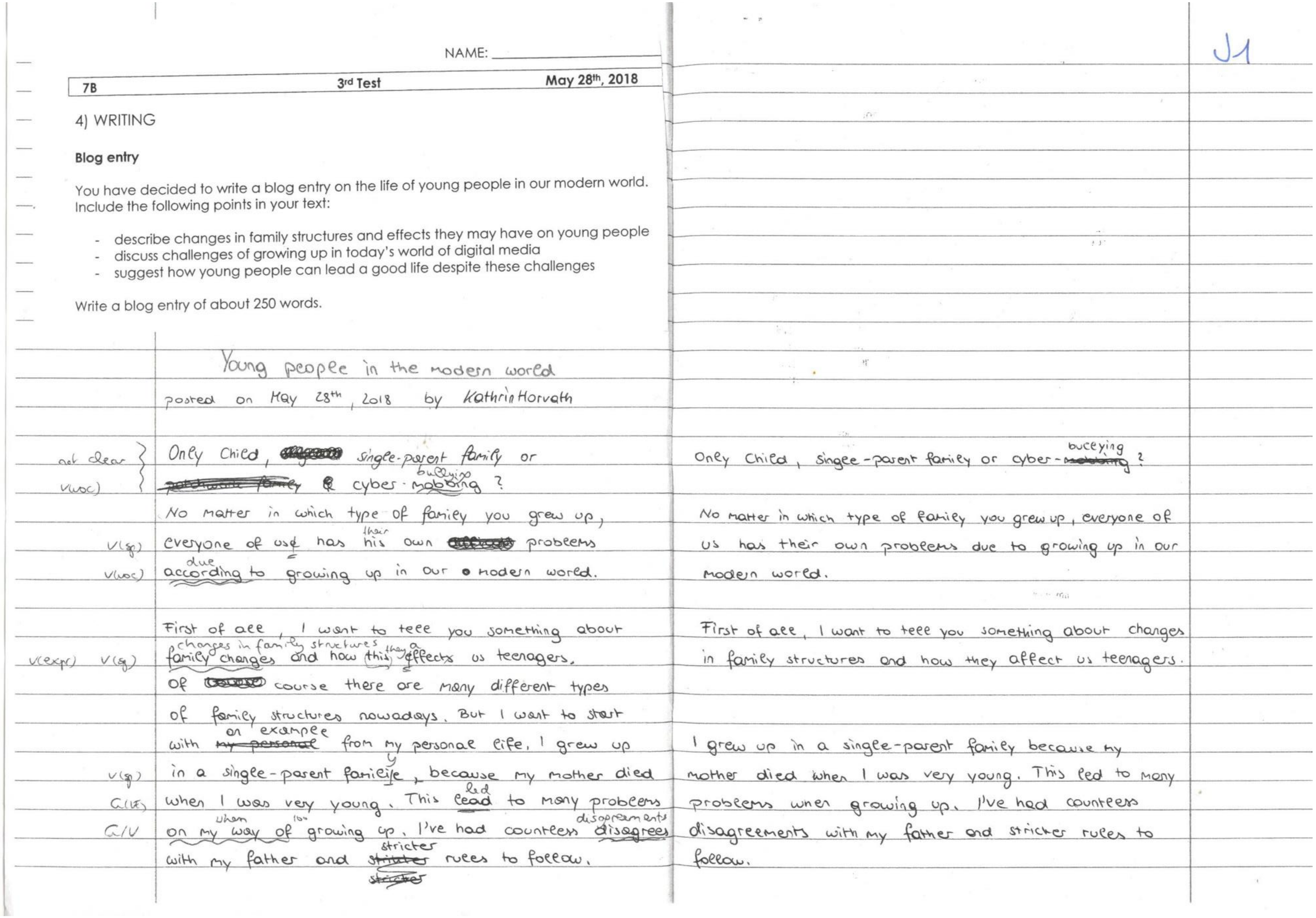
Task Achievement (TA)	6
Organisation & Layout (OL)	7
Lexical & Structural Range (LSR)	6
Lexical & Structural Accuracy (LSA)	6
TOTAL (max. 40 points)	25

m W 1st English examination	Nov. 21st, 2k18
The importance of ~~R~~ Reality TV	✓ Sp.
I was asked to write an essay about my opinion	Expr.
on Reality TV shows and the effect they have on	
the community. It's a fact that their importance	
is getting bigge and bigger. But why?	
Firstly, a lot of people watch them to focus on	Expr., Expr.
of people	
the lives ~~of~~ from another person with a different	WF, Prep., WW
lifestyle - maybe to escape their own lives a little	P
to	
bit and ~~y~~ get away from their daily struggles.	Λ, WF
The next point (I am going to talk about) is the problem	Expr.
with ~~R~~ Reality TV shows and how they may influence (the)	Sp., Expr
100	
society. The question you should ask yourself when	
~~a wat~~ watching them is; "How real they are at	Expr., P, WO
all?" Most parts of reality shows are scripted and	P
to attract the	
edited to be more interesting and achieve more	Λ, Expr.
audience might jealous of	
views. Some people maybe get mad or jelious on	WW, Sp., Prep.
the protagonists ~~and~~ because they have a good	WF
140	
live and a lot of money.	
The next thing is the viewers' influence on the	WF
shows. They want them to be more thrilling, dramatic,	
funny and sad at the same time and the producers	

try their best to make them watch their
show, so they often do something unexpected.
expectations get
Expr., WW, WF — But if (the) expectings getting higher and higher, there
when
WW, WF — will be a point (the) shows can't get better or
more exciting. At this point, the best way to
produce
WW — make a great show would be a real anymo
count
P, Λ — "Trueman Show." At this point, it doesn't ~~care~~ if
~~y~~ they use a single live ~~to~~ ~~[illegible]~~ and
play their game with it. The only thing that will be
audience figu
important is the entertainment and the views.

can
T — To sum up, I think reality shows could be
they get as w
Λ, WW, Λ — cool sometimes, but can (be) boring very fast
should not go
WF!, Sp. — But ~~I~~ they don't should go too far ~~just~~ only ~~to~~
for a good show.

Teacher J's prompt for the writing task in an exam of the 7th grade & the graded exam texts, J1 and J2:



J1

NAME: ____________

7B	3rd Test	May 28th, 2018

4) WRITING

Blog entry

You have decided to write a blog entry on the life of young people in our modern world. Include the following points in your text:

- describe changes in family structures and effects they may have on young people
- discuss challenges of growing up in today's world of digital media
- suggest how young people can lead a good life despite these challenges

Write a blog entry of about 250 words.

Young peopee in the modern wored

posted on May 28th, 2018 by Kathrin Horvath

not clear } Onely Chieed, single-parent family or
V(voc) ~~patchwork family~~ & cyber-~~mobbing~~ bullying ?

No matter in which type of familey you grew up,
V(sp) everyone of use has ~~his~~ their own problems
V(voc) ~~according to~~ due growing up in our modern wored.

First of aee, I want to teee you something about
V(expr) V(sp) ~~family changes~~ changes in family structures and how ~~this~~ they affects us teenagers.
Of course there are many different types
of famiey structures nowadays. But I want to start
with ~~my personal~~ an exampee from my personae eife. I grew up
V(sp) in a single-parent familie, because my mother died
G(LE) when I was very young. This ~~lead~~ led to many probeems
G/V ~~on my way of~~ when growing up. I've had counteess ~~disagrees~~ disagreements
with my father and ~~stritter~~ stricter ~~stricter~~ rules to foeeow.

Onely Chieed, singee-parent famiey or cyber-~~mobbing~~ bueeying ?

No matter in which type of famiey you grew up, everyone of us has their own probeems due to growing up in our modern wored.

First of aee, I want to teee you something about changes in famiey structures and how they affect us teenagers.

I grew up in a singee-parent famiey because my mother died when I was very young. This eed to many probeems when growing up. I've had counteess disagreements with my father and stricter rules to foeeow.

J1

But in the end we always discussed our probeems
and everything was fine.
Another exampee woued be my aunt, who was
V(voc) divorced and married her second ~~man~~ husband man, who had two chiedren.
C(t) So they form formed a patchwork famiey and at the
entence V(sp) beginning there where several probeems, Especiaeey for
the two kids, who had to accept another ~~women~~ woman
C(prep) V(voc) V(sp) besides dispite ~~of~~ her own mother a ~~caring for them~~, caring for them,

Secondey, there are aeso other chaeeenges for
(V(voc)) teenagers nowadays. Due According to the wored
V(sp) (V(voc)) of digitae media, cyber bullying mobbing is a big propleem 6 200
for young peopee. Chiedren and teenagers get
bueeied because of pics or videos
they post on several sociae networks. It's very
difficuet for young peopee to decide what kind
of photos and statements are ~~adequate~~ ~~right~~ ~~approper~~ adequate to post.

V(voc) Aee in aee, I want like you to know that no matter
(V(voc)) what chaeeenges you have to deae with, due according to
famiey life or digitae media, taek to your parents or
someone you trust! They ~~could~~ can help you
with famiey probeems and aeso ~~help you to deae~~
(V(voc)) with cyber-bullying mobbing.

C(vt) I am eooking forward to reading read your comments on
that topic. ~290 w

xoxo,
Kathrin

1) 12 r - 16,67%
2) 13 r - 19,12% } 71,57%
3) 26 r - 20,31%
4) 7,9,7,5 } 75,0%

Another exampee woued be my aunt, who was divorced
and married her second husband, who had two
chiedren. So they form a patchwork famiey and at the
beginning there were several probeems, especiaeey for
the two kids, who had to accept another woman
besides her own mother caring for them.

Due to the wored of digitae media, cyber bueeying is
a big probeem for young peopee.

Aee in aee, I want you to know that no matter what
chaeeenges you have to deae with, due to famiey
eife or digitae media, taek to your parents or someone
you trust!

I am eooking forward to reading your comments on that
topic.

Wassup, people

Have you ever thought about how family structures have changed because of digitalization? It recently came to my mind, so I thought, why not write a blog entry about it?

Nowadys, we can list a variety of types of families.

The nucleor family is still there, of course, but it has developed into many other types, as mentioned. Another current topic would be gay marriage. Long ago that wasn't allowed, but now it's a much-discussed topic. In my opinion, these changes in family structures don't affect the young generations in any way because literally anybody with decent intelligence can raise a child. It could happen that a child gets

J2

Our digital world

T: only part of the task

by NoLife69, May 28th, 2018, 11:17

Wassup, people!?

O(punct)

Have you ever thought about how family lives and structures have changed regarding due to/because of digitalization? It ~~so~~ recently came to my mind, so I thought, why not write a blog entry about it?

v(voc)

Nowadays ~~Today~~, we can list a ~~very~~ variety of types of families. There are patchwork families, blended and childless families, to give (you) some examples. The nuclear family is still there, of course, but it has developed into many other types, like as mentioned. A another current topic would be gay marriage (as well). Long ago, that wouldn't wasn't be allowed, but now it's a well much-discussed topic. In my opinion, these changes in family structures don't affect the new young generation(s) in any way, because literally anybody with decent intelligence can raise a child without it having problems. It could happen that a

patchwork = blended

a(prep)

a(v)

v(voc)

properly

108

raised isolated, but that is the parents fault then.
Not only families have changed, technology has too. Due to digitalization, digital devices are much-used nowadays. The younger generations, are very often using smartphones because they're very handy and almost everyone uses them. But that could also have negative consequences when we look at growing up. For example~~s~~, if a teenager doesn't possess a mobile phone, his social status will deteriorate instantly.

Well, you don't necessarily need digital devices.

It's your own choice, you can start a movement when and the way you want.

I hope I could motivate you a bit and maybe inspire you with my opinion

√2

child gets raised isolated, but that is the on the parents' fault parents then.
Not only families have changed, technology has did ~~it~~ too. Due to ~~the~~ digitalization, digital devices are much well-used nowadays. The younger newer generations, are very often ~~specialised~~ using specialised on smart-phones because they're very handy and almost everyone uses is using them. But that could also have mean negative consequences aspects when we look at growing up. For example, if a teenager doesn't possess a mobile phone, his social status abilieties will deteriorate (get worse) instantly too. Due to social media, everyone can access ~~others~~ profiles and data from others. Not being able to do that could lead to not being accepted.
Well, you don't necessarily need digital devices. They're helpful, but not compulsory. It's creativity and persuasive actions that make us independent. Don't be a sheep, don't let others say what you have to do. It's your own choice, you can start a movement when and the way how you want. You don't have to change your personality to be accepted, just be yourself.

I hope I could motivate you a bit and maybe get inspire you inspired with by my opinion on these

? not clear
v(expr)
G(t)
v(expr)
v(expr)
v(sp)
v(sp) G

J2

on these topics

topics. Feel free to add your own opinion in the comments, I will read them and respond to them. O(punct) corr

Thanks for reading this,

325 4

Gerald

[redacted]

4) 61 81 616 = 26,0 - 16,25%

[redacted]

Teacher A’s prompt for the writing task in an exam of the 7th grade:

Writing: Opinion Essay

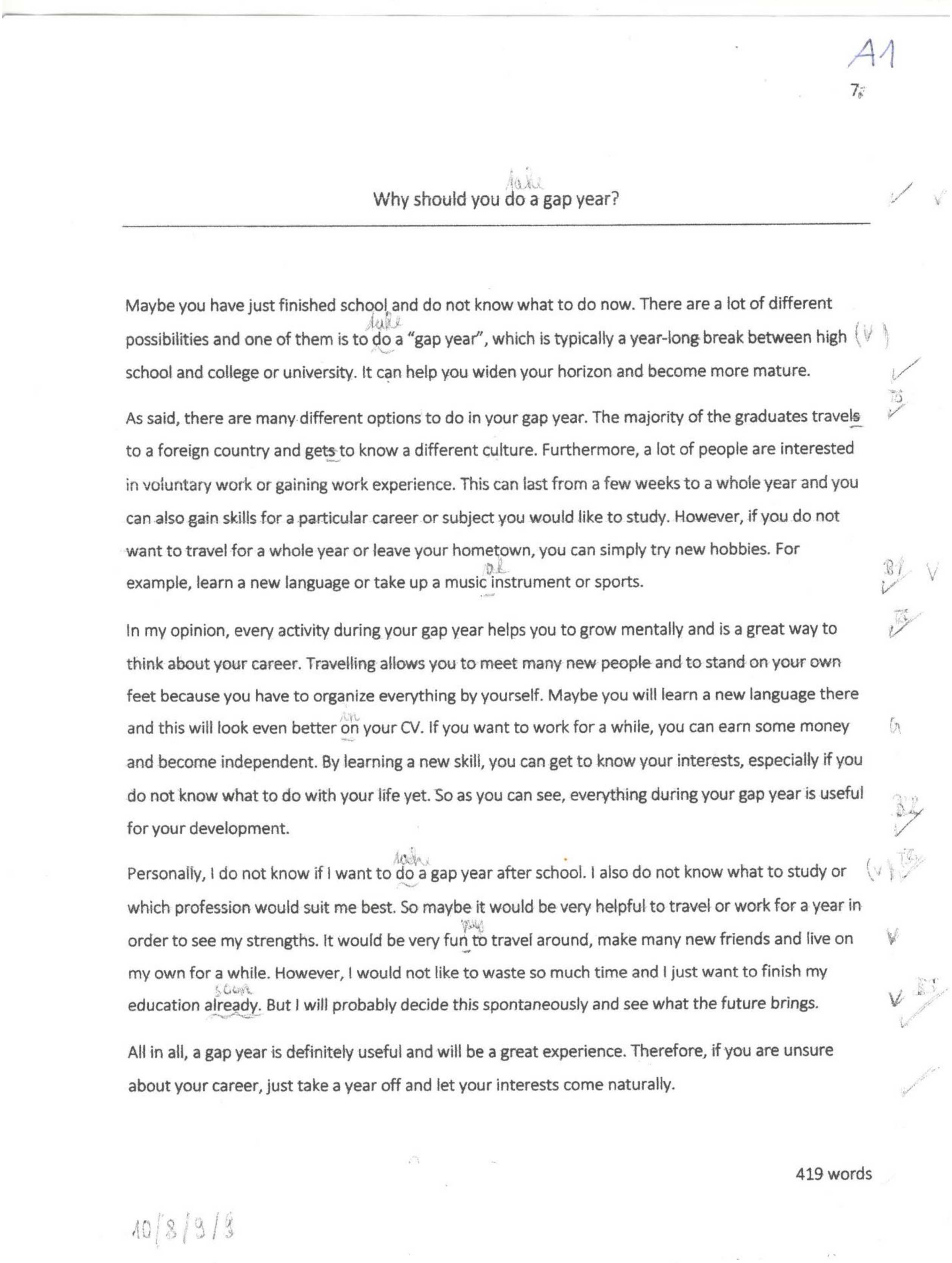
The Overseas Volunteer Organisation is running an essay competition with the title:
‘A Gap Year: Valuable Experience or Waste of Time?’

You decide to enter the competition. In your essay you should:

- explain the options that are open to school leavers in Austria.
- evaluate how useful each option is for the development of young people.
- comment on what you plan to do after school and why.

Write about 400 words.

Teacher A's graded exam texts, A1 and A2:



A1

7

Why should you do a gap year?

Maybe you have just finished school and do not know what to do now. There are a lot of different possibilities and one of them is to do a "gap year", which is typically a year-long break between high school and college or university. It can help you widen your horizon and become more mature.

As said, there are many different options to do in your gap year. The majority of the graduates travels to a foreign country and gets to know a different culture. Furthermore, a lot of people are interested in voluntary work or gaining work experience. This can last from a few weeks to a whole year and you can also gain skills for a particular career or subject you would like to study. However, if you do not want to travel for a whole year or leave your hometown, you can simply try new hobbies. For example, learn a new language or take up a music instrument or sports.

In my opinion, every activity during your gap year helps you to grow mentally and is a great way to think about your career. Travelling allows you to meet many new people and to stand on your own feet because you have to organize everything by yourself. Maybe you will learn a new language there and this will look even better on your CV. If you want to work for a while, you can earn some money and become independent. By learning a new skill, you can get to know your interests, especially if you do not know what to do with your life yet. So as you can see, everything during your gap year is useful for your development.

Personally, I do not know if I want to do a gap year after school. I also do not know what to study or which profession would suit me best. So maybe it would be very helpful to travel or work for a year in order to see my strengths. It would be very fun to travel around, make many new friends and live on my own for a while. However, I would not like to waste so much time and I just want to finish my education already. But I will probably decide this spontaneously and see what the future brings.

All in all, a gap year is definitely useful and will be a great experience. Therefore, if you are unsure about your career, just take a year off and let your interests come naturally.

419 words

10/8/9/9

A2

7.

„A Gap Year: Valuable Experience or Waste of Time?"

→ do not copy from task

take a stand / state your opinion

Graduation is an exciting time. It is both an ending and a beginning, it is memories of the past and big dreams for the future. But what to do afterwards? I am convinced doing (taking) a gap year is never a mistake. ✓

First of all, there are several possibilities for an Austrian graduate to do (choose from) after finishing school. On the one hand, you are able to make (take) a social year or gap year, where you attend a charity for example, and try to find the right path for your upcoming life. On the other hand, you can just start with attending university and with studying medicine or law for instance.

You may not find the right path with your first try, but regardless of what you do after graduating, it will be useful for your development and your entire life. Furthermore, a gap year is often a once in a lifetime experience and the chance to escape the daily grind. However, if planned right, it will also be an educational opportunity of growth and other benefits and not just a "vacation" or year off. Moreover, if you start attending university directly after your graduation, you do not waste any time and you will finish your studies earlier than the other ones.

Finally, I want to tell you what my plans after graduation are. Actually, I would love to make (take) a gap year, but I want to study medicine, and this takes about seven to nine years to finish. So, as you can imagine, I do not want to squander any valuable time I can use for my studies. If it is possible, I would ~~very much~~ love to make (take) a few "gap months" after graduating from university and travel the world (or just a few countries) alongside some friends or my family. After all, nothing is fixed and unfortunately, nothing in life works how you plan and wish for it.

All in all, your work and life after school and college are going to fill a large part of your life, and the only way to be truly satisfied is to do what you believe is great work. And the only way to do great work is to love what you do. So, I believe that doing (taking) a gap year is a very good opportunity to get on the right track but if you are not interested in it, you should do whatever you want to do, or you want to become.

words: 416

9/10/9/8

9.4. Zusammenfassung

Die Einführung der Zentralmatura in Österreich hat eine Veränderung in vielen Bereichen mit sich gebracht, so auch im Bereich des Beurteilens der Prüfungen. Im Bereich des Schreibens im Unterrichtsfach Englisch bedeutete dies eine Standardisierung der Beurteilung und der Einführung eins analytischen Beurteilungsrasters. Der Fokus dieser Arbeit liegt auf dem Umgang der Lehrer und Lehrerinnen mit diesem Instrument im Zusammenhang mit Gerechtigkeit, Objektivität und Voreingenommenheit. Um Informationen über die Lehrer und Lehrerinnen und deren Handhabung des Beurteilungsrasters zu erhalten, wurde eine Recherche angestellt, die aus Fragebögen und Einsicht in die Korrekturgewohnheiten der teilnehmenden Lehrpersonen bestand. Die Informationen gaben Aufschluss darüber, dass das Beurteilen von Texten nie objektiv und unvoreingenommen passiert und persönliche Präferenzen eine Rolle spielen können, über genaue Einflüsse konnte jedoch keine Angabe gemacht werden.

9.5. Abstract

The implementation of a new standardised school-leaving exam brought many changes to the Austrian EFL classroom, especially in relation to assessment. A new assessment tool was created by BIFIE to rate students' text in the exam with an analytic rating scale that distinguishes four rating criteria. This paper focuses on the teachers who grade the texts and their grading process in connection with fairness, objectivity and bias. To obtain relevant information which can be used to analyse the connection between the rating process and the influence of the raters' personal preferences a qualitative research method was used by looking at students' (B1 & B2 level) essays which have been marked and graded with the CEFR Linked Austrian Assessment Scale (CLAAS) by teachers who underwent rater training and have been working with it. The research showed that rating errors occur and raters can be influenced by various factors.