colloquium:
new philologies

The Role of Errors in Validating a Large-Scale Assessment of Adolescent English Writing in Austria

Samuel Hafner
University of Klagenfurt (Austria)

Günther Sigott
University of Klagenfurt (Austria)

Abstract

This study investigates errors in a sample of 50 written performances of Austrian learners of English collected in the 2009 baseline study for the Austrian Educational Standards-Based Writing Test for English at grade 8 (E8 Standards Writing Test). The research aims to contribute to the validation of this large-scale assessment by studying the relationship between errors (described using the *Scope – Substance* error taxonomy) and human ratings awarded to writing performances. The results add to the validity evidence of the E8 Standards Writing Test. There is a negative relationship between human ratings and the presence of errors; a low error density is associated with higher ratings and a high error density with lower ratings. *Substance* WORD, CLAUSE, and TEXT error densities play an important role in the rating in most dimensions; errors with a larger *scope* also have a strong effect. By highlighting aspects of errors to which raters seem to be sensitive, these findings constitute evidence of context validity. At the same time, the findings are relevant to theory-based validity by concretising areas of competence that learners need to develop in order to receive higher ratings. While errors are important determinants of the ratings, additional factors, presumably positive features, must be at play as the accuracy of the regression models is low to moderate. This should in fact be the case since the E8 rating scale refers to negative as well as positive features.

Keywords: *Scope – Substance error taxonomy*, *E8 Standards Writing Test*, *Austrian Educational Standards in English*, *Error analysis*, *Validation*, *Validity*, *Context validity*; *Theory-based validity*

Colloquium: New Philologies, Volume 6, Issue 2 (2021)
doi: 10.23963/cnp.2021.6.2.4
Stable URL: https://colloquium.aau.at/index.php/Colloquium/article/view/143

1 Introduction

Analytic rating scales for writing, and the rating scale used in the Austrian Educational Standards-Based Writing Test for English at grade 8 (E8 Standards Writing Test) in particular, describe levels of performance on each of several dimensions by means of so-called performance level descriptors (e.g. Cizek and Bunch 2007, 46). These are supposed to guide trained raters in the process of judging the writing performances with the aim of maximising intra- and interrater agreement, which, in turn, is considered as an indication of validity. In fact, research into the validity of ratings tends to rely on rater agreement as an indication of validity. However, there is little evidence on the extent to which errors, which are only one aspect that is covered by the performance level descriptors in the rating scales, play a role in the raters' rating behaviour. More particularly, little is known about the extent to which the raters actually pay attention to errors, and about the kinds of error to which they are sensitive, when forming their judgments.

This study aims to shed light on this aspect of the rating process in the E8 Standards Writing Test. More specifically, this research aims to contribute to the validation of this test by investigating the role that errors play in the rating of writing performances using an analytic rating scale. The study identifies the most common errors in the writing of Austrian learners of English at the age of around 14 years and, in particular, focuses on the relationship between errors and human ratings awarded to writing performances generated in the context of the E8 Standards Writing Test. The errors in the writing performances are identified and categorised by means of the *Scope – Substance* error taxonomy (Dobrić and Sigott 2014).

In order to address the role of errors in the rating of the E8 Standards Writing Test, the following research questions are investigated:

RQ1. In terms of errors described using the *Scope – Substance* error taxonomy, what are the most common errors in texts written by Austrian learners of English at grade 8?

RQ2. What is the relationship between errors described using the *Scope – Substance* error taxonomy and the human ratings in the four dimensions of *Grammar*, *Vocabulary*, *Coherence & Cohesion*, and *Task Achievement*?

2 Theoretical Framework

2.1 Relation to Validation Theory

Studying the influence of errors on ratings of writing performances addresses aspects of validity. More concretely, this study investigates the relationship between the error incidence and the ratings which were awarded by trained raters using an analytic rating scale. By doing this, it aims to identify error types that are typical of performances with a lower rating. Avoiding these error types constitutes a challenge which test takers have to master. Avoiding these errors thus constitutes aspects of task difficulty. Identifying the kinds of errors that constitute such difficulty contributes to identifying the construct underlying the E8 Standards Writing Test. In fact, understanding difficulty-generating features in any test is central to understanding what the test tests. In the assessment of writing, difficulty-generating features stem from the test context, such as the task set, the rating criteria, and the raters. The rating criteria are an important source of difficulty-generating features. However, raters looking at the rating criteria per se is no guarantee that they will follow these criteria to the letter, either because they interpret them in their own way or because they develop their own rating strategies when they encounter problems in applying the criteria (Lumley 2005). Therefore, evidence is needed of what the raters actually do when they rate a performance. This evidence can be collected by trying to access the raters' cognitive processes during rating or by observing features of performances that tend to cooccur with individual levels awarded to these performances by the raters. These cooccurrences, or correlations, are empirical evidence which suggests what it is in the performances that raters actually pay attention to in the rating process. These correlations can, however, only be interpreted as suggestive evidence since they are statistical, not causal, relationships. Nevertheless, they indicate features which cooccur with particular ratings. It follows that changes in the occurrence of these features will go hand in hand with changes in the rating. Features correlated with low ratings will, when avoided, make higher ratings likely. The necessity of avoiding such errors, then, can be considered to be a difficulty-generating feature. Consequently, identifying such errors contributes to our understanding of what the test tests, or, put another way, to our ability to identify the test construct. This is why Sigott (2004, 51) has referred to this aspect of validation as *construct identification*.

In terms of Weir's (2005) validation framework, the study addresses *context validity* and *theory-based validity*. Context validity refers to the totality of contextual factors that influence the production of test-taker performance as well as to the criteria for correctness that are applied in scoring the performance. Raters' sensitivity to error types constitutes such expectations of correctness. Identifying the errors that raters attend to when rating therefore provides evidence of context validity. This also contributes to our understanding of rater cognition (Dobrić 2020). From the writer's perspective, avoiding such errors constitutes challenges, or sources of difficulty (Dobrić et al. 2021). Specifying these challenges means specifying aspects of the test construct and hence addresses aspects of theory-based validity.

2.2 Error Analysis

Error analysis (EA), as a branch of applied linguistics, is concerned with the study and analysis of errors made by L2 learners. Building on the work of Corder (1974), Ellis and Barkhuizen (2005, 57) distinguish the following five steps in conducting an EA: (1) collecting a sample of learner language, (2) identification of errors, (3) description of errors, (4) explanation of errors, and (5) error evaluation. Steps two and three are of interest in this study and will therefore be explained further.

The second step, *identification of errors*, involves the recognition of elements in the learner's production that deviate from the norm of the L2 in some way. According to Ellis and Barkhuizen (2005), identification "involves a comparison between what the learner has produced and what a native speaker counterpart would produce in the same context" (58). This description builds on the work of Lennon (1991), who defines error as "a linguistic form or combination of forms which, in the same context and under similar conditions of production, would, in all likelihood, not be produced by the speakers' native speaker counterparts" (182).

An important process here is to "prepare a reconstruction of the sample as this would have been produced by the learner's native speaker counterpart" (Ellis and Barkhuizen 2005, 58). The reconstructed version is referred to as *authoritative reconstruction* of the learner performance. Arriving at such a version is often not without problems. In many cases, the intended meaning of the sentence/utterance is not clear and thus, several reconstructions are possible. For such cases, Corder (1974) suggests seeking an *authoritative interpretation* by asking the learner what they intended to communicate (127–128). This procedure, however, is unpractical for most EA. Therefore, he proposes plausible interpretation as an alternative. Here, the researcher has to "attempt to infer the meaning intended by the learner from the surface structure and his text-sentence in conjunction with the information derived from its context" (128).

The third step, *description of errors*, involves "specifying how the forms produced by the learner differ from those produced by the learner's native speaker counterparts" (Ellis and Barkhuizen 2005, 60). Consequently, the most important process in this step is choosing or developing an error classification system with descriptive categories for coding the errors which have been identified. A system of error categories is referred to as *error taxonomy* (see, e.g., James [1998, ch. 4] for an overview of different approaches).

EA does have some methodological shortcomings and limitations. First, although learners make errors in both comprehension and production, EA can only deal effectively with errors in speaking and writing because errors only manifest themselves visibly in the productive skills. Second, EA gives an incomplete picture of the learning process because it focuses on what the learners do wrongly and not on what they already know

(Hammarberg 1974, 185; Ellis and Barkhuizen 2005, 70; Saville-Troike 2006, 40). Additionally, EA cannot account for learner use of communicative strategies such as avoidance of difficult structures. The absence of errors in a certain area of the language, thus, does not mean that the learner has mastered it (Khansir 2012, 1030; Saville-Troike 2006, 40; Schachter 1974; X. Yu 2017). Third, Lennon (2008) points out problems in assigning the psycholinguistic cause of errors because error explanation is in its nature speculative, and it is often not possible to unambiguously locate the source of an error. Last, there is ambiguity in error classification and identification (Ellis and Barkhuizen 2005, 59–60; Lennon 2008, 55; James 1998, 91–92). The *Scope – Substance* error taxonomy (Dobrić and Sigott 2014) was created as an attempt to alleviate this problem.

Although the research focus of EA has changed, it is still useful as a "methodology for dealing with data, rather than a theory of acquisition" (Cook 1993, 22). This is precisely how EA has been used in studying the impact of learner errors on human ratings of learner performances (e.g., Pibal 2012; Pibal, Sigott, and Cesnik 2018), and this is also the purpose for which it is used in the present study.

2.3 The *Scope – Substance* Error Taxonomy

The *Scope – Substance* error taxonomy is an error classification system that was proposed by Dobrić and Sigott (2014) to alleviate problems with subjectivity in existing taxonomies. In this model, errors are classified on the basis of two concepts: *substance* and *scope*. *Substance* is defined as the "the size of the element that needs to be changed in order to correct the error", while *scope* refers to "the amount of textual or extratextual context that is required for recognising the presence of an error" (114). The idea of characterising errors in this manner was originally proposed by Lennon (1991), as the authors of the taxonomy point out (Sigott, Cesnik, and Dobrić 2016, 80). Similar to the concepts of *scope* and *substance*, Lennon suggested classifying errors by means of their *extent* and *domain*.

The core idea of the *Scope – Substance* error taxonomy is that most errors only become identifiable when there is an incompatibility between a unit in the text (the potential error) and the surrounding context. As Lennon (1991) points out, "most 'erroneous forms' are, in fact, in themselves not erroneous at all, but become erroneous only in the context of the larger linguistic unit in which they occur" (189). For example, looking at the word 'neccessary' in isolation is enough to see that it is erroneous. However, in the clause 'I was very nervously', every word and phrase is in itself acceptable. Only when we consider the context of the whole clause does it become apparent that the use of the adverb 'nervously' is not correct.

The size of the constituent in which the error is located and the surrounding context that needs to be checked for incompatibility is expressed in terms of syntactic units. These are described in terms of units of the commonly accepted grammatical hierarchy, starting with morpheme, the smallest unit, which makes up words, which make up phrases, which make up clauses, which make up sentences. These units are placed in a hierarchy on the basis of their potential size or extensibility and not their actual size or length (referring to the number of constituents) as a sentence may consist of only one clause, or a phrase of only one word (e.g., all three phrases in 'I am nervous'). The phenomenon that one unit may be the only constituent into which another unit can be analyzed is referred to as unitary constituency (Quirk et al. 1985). In the most extreme case, for instance 'Bye.', a structure may be described as a sentence, a clause, a phrase, a word, or a morpheme. The *Scope – Substance* error taxonomy uses the principles of syntactic analysis laid out in Quirk et al. (1985) since a theory-neutral, descriptive grammar is very well suited as a framework for error description (cf. James 1998, 96). In contrast to other, more recent, grammars of English, it is likely that almost every English sentence structure, even learner language, is describable following its principles.

In the taxonomy, errors are described by a combination of *scope* and *substance*. Using the units from the grammatical hierarchy without morpheme but with punctuation and text added, the latest version of the taxonomy has 19 different error types (Sigott, Cesnik, and Dobrić 2016) (see Table 1).

Code	***Substance***	***Scope***	**Code**	***Substance***	***Scope***
11	WORD	WORD	34	CLAUSE	SENTENCE
12	WORD	PHRASE	35	CLAUSE	TEXT
13	WORD	CLAUSE	44	SENTENCE	SENTENCE
14	WORD	SENTENCE	45	SENTENCE	TEXT
15	WORD	TEXT	55	TEXT	TEXT
22	PHRASE	PHRASE	92	PUNCTUATION	PHRASE
23	PHRASE	CLAUSE	93	PUNCTUATION	CLAUSE
24	PHRASE	SENTENCE	94	PUNCTUATION	SENTENCE
25	PHRASE	TEXT	95	PUNCTUATION	TEXT
33	CLAUSE	CLAUSE			

Table 1: The 19 error types of the updated taxonomy (with error codes).

Mathematically, 30 combinations would be possible. However, the *scope* of an error cannot be lower than its *substance*, which eliminates 10 combinations like *substance* PHRASE – *scope* WORD. Additionally, *substance* PUNCTUATION – *scope* WORD is not a possible error type (Sigott, Cesnik, and Dobrić 2016).

In Figure 1, the different error types are represented diagrammatically using a coordinate system, showing graphically that the *scope* of an error cannot be lower than its *substance*.

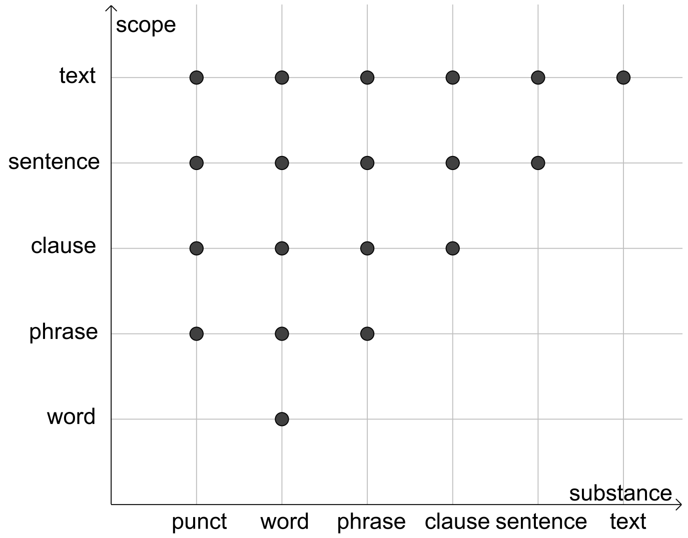


Figure 1: Diagrammatic representation of the 19 error types.

Following are examples that illustrate the application of the taxonomy (for additional examples see Dobrić and Sigott 2014; Sigott, Cesnik, and Dobrić 2016; Hafner 2018a).

***substance* word – *scope* word**

(1) If I could fly, I would go to Madrid and get an autogramm from Ronaldo.

In the first example, the word 'autogramm' is not part of the English lexicon. In order to repair the error, a correction on the word level must be performed, making the *substance* WORD. The *scope* of the error is WORD since it is sufficient to look at the word in isolation to identify the error.

***substance* word – *scope* text**

(2) We go around with them, talk with them, and have a lot of fun.

This sentence does not seem to have any erroneous forms at all when looked at in isolation. However, if the context is widened to the text level, it becomes clear that the learner is talking about events in the past. Therefore, the verbs must be put into past tense. These errors have *substance* WORD and *scope* TEXT.

***substance* phrase – *scope* clause**

(3) When you will phone me, I'll come.

The tense of the verb phrase 'will phone' is not acceptable in a clause that is introduced with 'when'. Hence, the auxiliary verb 'will' needs to be deleted, which changes the structure of the verb phrase (from having an auxiliary slot to not having one). The error becomes visible when the *scope* is widened to the clause level ('When you will phone me'). Consequently, the *substance* is PHRASE while the *scope* is CLAUSE.

***substance* TEXT – *scope* TEXT**

This error type typically occurs when the conventions of text structure are violated, such as the logical order of sentences (topic sentence – supporting details – concluding sentence; cause-effect relations, etc.).

(4) Ten people are on vacation on a boat. The star is Ben Klin Hof. A shark kills one of the people while he was swimming in the sea.

This extract is from a paragraph describing the plot of a movie. In isolation, each of the sentences is acceptable. However, the second one is not appropriate in this slot. Correcting the error involves changing the position of this sentence into a more appropriate one, which changes the text structure. Therefore, both *substance* and *scope* are TEXT.

***substance* PUNCTUATION – *scope* SENTENCE**

For the sake of reproducibility, only obligatory commas are considered errors (e.g., lists, initial adverbials that are clauses, non-defining relative clauses, appositions, question tags, introductory comma in informal letters).

(5) When we arrived there I was very sad.

Since the temporal clause 'When we arrived there' comes before the main clause, the comma between the two clauses is obligatory. The correction thus involves inserting a comma, making the *substance* PUNCTUATION. The error becomes apparent when the whole sentence is considered, thus the *scope* is SENTENCE.

2.4 Austrian Educational Standards-Based Writing Test for English at Grade 8 (E8 Standards Writing Test)

The E8 Standards Writing Test is part of a large-scale secondary-level monitoring program of EFL competences at grade 8 in Austria. The main focus of the program is to monitor the listening, reading, writing, and with a restricted sample, speaking abilities in EFL of pupils in grade 8 and to provide diagnostic information to help transform the teaching and learning of EFL in the Austrian school system (Kulmhofer and Siller 2018). The test takers are approximately 14 years old when they take the E8 Standards Writing Test and should have reached the CEFR levels A2 to B1, depending on the area of competence and descriptor. In 2009, a baseline study was conducted, the first nation-wide Standards Test was administered in 2013, and the most recent one took place in 2019.

In the E8 Standards Writing Test, candidates are assessed in four areas: *Task Achievement*, *Coherence & Cohesion*, *Grammar*, and *Vocabulary*. These areas represent the four dimensions of the analytic rating scale used for the assessment of the writing performances. Each dimension has seven levels plus a '0' band, which is reserved for performances with no assessable language. Four levels (1, 3, 5, 7) have descriptors, while the three levels in between (2, 4, 6) are empty (Gassner, Mewald, and Sigott 2008). The rating scale used for the 2009 baseline study (version May 2008) is reprinted in the appendix.

In the 2009 baseline study, each student received a booklet containing a short and a long answer prompt. The candidates had 10 minutes to produce the short performance and 25 minutes for the long one; time management, however, was left to the students. The prompt instructed the students on the required length of the text and the content points to be addressed. The short answer prompt demanded between 40–60 words and contained between three and five bullet points, and the long answer prompt asked for 120–150 words with five to eight bullet points (Kulmhofer and Siller 2018). Additionally, the writing prompts specified language functions like inviting, apologizing, asking for something or giving advice that the test takers need to perform (Gassner et al. 2011).

3 Research Design

3.1 Sample and Population

The sample of written performances came from the 2009 baseline study for the E8 Standards Writing Test, in which a stratified random sample of 10,749 eighth-grade pupils from 204 schools from all over Austria participated (IQS n.d.). The performances were taken from an already existing corpus compiled by Pibal (2012), who selected a random sample of 100 long performances from the overall pool. Pibal only selected long performances (120–150 words in length) because they offer more potential for analysis than the short ones. The texts are based on two different prompts. One instructed the students to write a letter to a friend or relative in which they talk about a recent school trip they undertook. The other required pupils to write a text for a youth magazine and describe their favourite film, music, or book.

Pibal digitized the texts for the purpose of further analysis by transcribing them into text files. From this corpus, a random set of 50 performances was drawn for this research. Each text was assigned a unique number from 1 to 50. Two texts had to be excluded (no. 43 and 45) from the study as they proved to be unsuitable for analysis. The performances were incoherent, full of errors, and the writers' intended meaning was unclear.

3.2 Variables

3.2.1 Adjusted Human Ratings (Fair Measures)

After the administration of the E8 Standards Writing Test in 2009, all performances were rated by practising EFL teachers from Austrian schools, who went through a specific rater training program, using a seven-band analytic rating scale (see appendix). In order to be able to minimize the effect of differences in rater behaviour, a rating plan with rater overlap for multiple rating of anchor performances was followed. The raw ratings were then subjected to multifaceted Rasch analysis (MFRA) to adjust for differences in rater severity and task difficulty. The output of this analysis is called *Fair Measure*. A person score based on the aggregated ratings of two performances (long and short) was reported for each participant. Pibal (2012) re-conducted MFRA of the original data on a single-performance basis and calculated five Fair Measure variables: Total Fair Measure, Grammar Fair Measure, Vocabulary Fair Measure, Coherence & Cohesion Fair Measure, and Task Achievement Fair Measure. As the raters did not award an overall rating, the Total Fair Measure is an artefact, aggregated from the Fair Measures on the four rating dimensions. Pibal's single performance-based Fair Measures were used in the present study. Table 2 shows how the variables are named in the analysis. The Fair Measures can take a value between 0 and 7, with 7 being the best rating.

Variable	Variable name
Total Fair Measure	TOT.Fair
Grammar Fair Measure	GR.Fair
Vocabulary Fair Measure	VOC.Fair
Coherence & Cohesion Fair Measure	COH.Fair
Task Achievement Fair Measure	TA.Fair

Table 2: Adjusted human rating (Fair Measure) variables

3.2.2 Error Density (ED)

To make the occurrence of errors comparable across performances and relating the errors to the human ratings, error density variables were calculated. The error density is defined as the number of errors per one hundred words. Table 3 gives an overview of the variables, their naming pattern, and their computation.

This was done for two reasons. First, the longer the text, the more potential for error it offers. Second, raters usually take the whole performance as a basis for their judgment. Thus, the same number of errors in a short text will, most likely, be judged more severely than in a longer one.

The error density variables take this into account. For example, the total error density (ED_TOT) for a 100-word text with 10 errors is 10; for a 200-word text with the same number of errors, the value is 5. Consequently, the error density variables allow a more informative comparison of the occurrence of errors independently of the text length, which would not be possible with the absolute frequencies.

Level	**Error density variables**		**Computation**
total	total error density (ED_TOT)	(1 variable)	$\dfrac{\text{total number of errors}}{\text{number of words}} \cdot 100$
substance	*substance* X density (ED_SuX)	(6 variables)	$\dfrac{\text{number of errors with } \textit{substance} \text{ X}}{\text{number of words}} \cdot 100$
scope	*scope* Y density (ED_ScY)	(5 variables)	$\dfrac{\text{number of errors with } \textit{scope} \text{ Y}}{\text{number of words}} \cdot 100$

ED $=$ error density. X $\in \{1, 2, 3, 4, 5, 9\}$ and Y $\in \{1, 2, 3, 4, 5\}$.
$1 \hat{=}$ WORD, $2 \hat{=}$ PHRASE, $3 \hat{=}$ CLAUSE, $4 \hat{=}$ SENTENCE, $5 \hat{=}$ TEXT, $9 \hat{=}$ PUNCTUATION.

Table 3: Overview of the error density (ED) variables, their naming pattern, and their computation.

3.3 Guidelines and Agreement Study

The quality of the linguistic information in any annotated learner corpus depends on the reliability of the error tagging (Díaz-Negrillo and Fernández-Domínguez 2006, 88; Dobrić 2015, 36; Plaban, Pabitra, and Anupam 2008, 58). More generally, the validity of a research study is strongly dependent on the reliability of the data (Brants 2000; Plaban, Pabitra, and Anupam 2008, 58). Granger (2003, 467) describes consistency as one of the key requirements for an annotation system to be fully effective. To achieve consistency, she recommends the elaboration of an error manual with detailed descriptions of the error categories and tagging principles (466–467). A similar point was made by Potter and

Levine-Donnerstein (1999), who emphasize the need for a coding scheme, that is, a set of rules that tell annotators how to code (266–267). Using such a manual or coding scheme should help to ensure that the error coding is the result of a systematic examination with minimal uncertainty and subjectivity which should, as a result, lead to higher levels of reliability. Taking this into consideration, an elaborate set of rules and principles (collectively referred to as guidelines) was developed. These guidelines are the product of a collaborative process involving a group of researchers working with the *Scope – Substance* error taxonomy (Bıdık 2016; Hafner 2018b; Sigott, Cesnik, and Dobrić 2016; Steinkellner 2018) and including an agreement study. For more details on this process and the full guidelines, see Hafner (2018a).

3.4 Data analysis

The study utilizes a mixed methods approach. The qualitative aspect is the identification and description of errors in the performances. Correlation and regression analyses represent the quantitative part.

3.4.1 Identification and Description of Errors

Building on the theoretical foundation laid out by Ellis and Barkhuizen (2005), the following three-step procedure was applied to each written performance.

- First, the whole text was read to develop an overall understanding.
- Then, the process of error identification started. This process was stepwise and bottom-up, starting with the smallest unit in the grammatical hierarchy. This means that for every sentence in a performance, the individual words were first screened for errors. The process was then repeated for the phrase, clause, sentence, and text level. The output of this procedure was an authoritative reconstruction of the learner language. The guiding principle was to reconstruct the intended meaning of the learner (plausible interpretation) while changing as little as possible (minimal correction).
- Last, the process of error description started. The errors were coded with ATLAS.ti (version 8). After importing the digitized texts into the software, the error type codes from Table 1 were used to create 19 code labels (e.g., 11, 12, 13, ...) each indicating a combination of *substance* and *scope*. These formed the basis for error tagging. Each data segment constituting an error *substance* was marked with one of these codes. The coding was performed by one of the authors [SH], after consultation with other researchers [GS and FS] also working with the taxonomy, in cases of uncertainty. The annotator analysed each text at least twice. To

ensure reliability and consistency of the application of the *Scope – Substance* error taxonomy, an agreement study was carried out before the annotation process (see section 3.3).

3.4.2 Statistical Analyses

After the categorization process in ATLAS.ti 8, the data file was merged with the ratings provided by Pibal (Pibal 2012). The statistical analysis was performed with the statistical software R (version 4.1.0).

Descriptive statistics were used to describe the basic features of the data in the study and to provide simple summaries about the sample. The main aim here was to give an insight into the areas that cause the most difficulty for Austrian learners of English at grade 8 (RQ1).

RQ2 was answered using correlation and regression analysis. To assess the strength of the (statistical) association, Spearman's rank-order correlation coefficient (Spearman's ρ) was calculated as not all variables are normally distributed and outliers are present. Pearson's product-moment correlation coefficient (Pearson's r) is reported too for comparison as ρ benchmarks *monotonic* relationships, while r assesses *linear* ones. In linear relationships, variables tend to move together at a constant, i.e., linear, rate, while in a monotonic relationship, variables tend to move in the same relative direction, but not necessarily at a constant rate.

In short, correlation quantifies the degree to which two variables are related. In contrast, linear regression analysis provides information about the change in the dependent variable when the values in the independent variable(s) change. A simple linear regression model is a mathematical equation that allows us to predict the value of Y (dependent variable) for a given value X of the independent variable:

$$Y = b + a \cdot X$$

Thus, regression is used for the mathematical modelling of the relationships between the Fair Measures (dependent variables) and the error densities (independent variables). For this study, we conducted *descriptive* modelling to analyse the impact of the independent variables on the dependent ones without assuming or relying on an underlying causal theory.[1] Since the error densities have skewed distributions and outliers are present in the data, the regression coefficients are estimated using an MM-Estimator, a highly robust and highly efficient estimator (R. R. Wilcox 2012, 499), as implemented in the function `lmrob` in the R-package `robustbase` (version 0.93-7) (Maechler et al. 2021) with

[1] See (Shmueli 2010) for an excellent overview of the different kinds of statistical modelling.

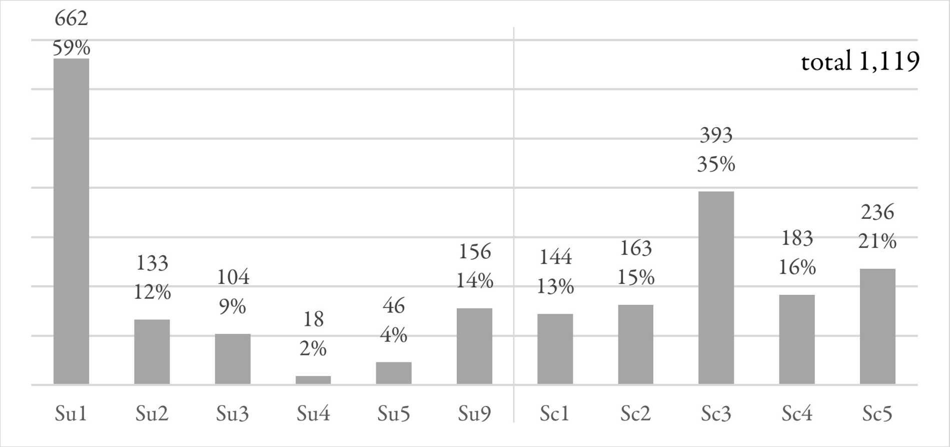
the option `setting="KS2014"` (cf. Koller and Stahel (2011, 2017) and the reference manual of the package on the Comprehensive R Archive Network (CRAN) website). All models were tested for homoscedasticity of residuals (equal variance), autocorrelation, multi-collinearity, and normality of residuals.

4 Results

4.1 Descriptive Statistics

In total, 1,119 errors were tagged across all 48 texts. On average, every performance contains about 23 errors with a standard deviation of about 10. The median is 22. Virtually all texts contain between 7 and 41 errors and 50% of the performances have between 16 (lower quartile Q1) and 28 errors (upper quartile Q3).

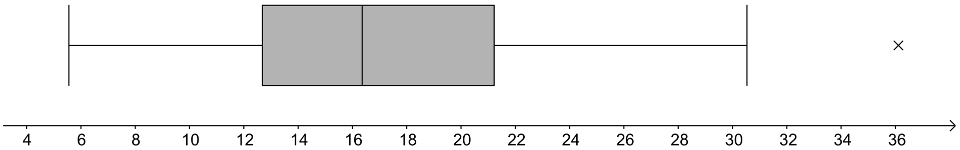
In order to answer RQ1 (*In terms of errors described using the Scope – Substance error taxonomy, what are the most common errors in texts written by Austrian learners of English at grade 8?*), the frequencies of the different *scope* and *substance* errors are shown in Figure 2. *Substance* word errors are by far the most frequent (59% of all errors). Larger *substance* errors (clause, sentence, text) only make up 15% of all errors. Errors with *substance* punctuation represent the second most common *substance* group (14% of all errors). Errors with a *larger scope* (clause, sentence, text) constitute almost three quarters (73%) of all errors. *Scope* clause errors are the most frequent errors among all *scope* categories and constitute 35% of the errors, followed by *scope* text (21%).



SuX = all errors with *substance* X, ScY = all errors with *scope* Y. 1 ≙ word, 2 ≙ phrase, 3 ≙ clause, 4 ≙ sentence, 5 ≙ text, 9 ≙ punctuation

Figure 2: Error frequencies (*scope*/*substance* categories).

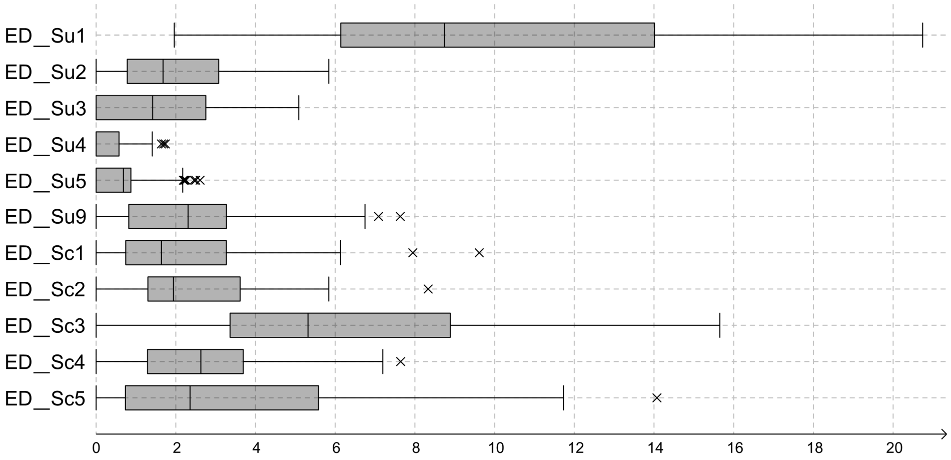
On average, the performances contain 17.03 errors per 100 words with a standard deviation of 6.96. The box plot in Figure 3 displays the distribution of the total error density. The diagram shows that the total error density is approximately normally distributed. In 50% of the performances, the value lies between 12.67 (Q1) and 21.41 (Q3). 50% of the texts have a value of more than 16.36 (median). The minimum is 5.56 and the maximum 36.11. However, the total error density of 36.11 is an extreme value and virtually all texts have between 5.56 and 30.53 errors per 100 words.



N = 48, min = 5.556, max = 36.111, Q1 = 12.672, median = 16.359, Q3 = 21.412. The scale refers to number of errors per 100 words. The whiskers have a maximum length of 1.5 times the interquartile range.

Figure 3: Box plot of total error density (ED_TOT).

The box plots in Figure 4 give more insight into the distribution of the *scope* and *substance* error densities. None of these variables is normally distributed. The tails of the



The scale refers to number of errors (of a specific type) per 100 words. The whiskers have a maximum length of 1.5 times the interquartile range. The extreme value for 'ED_Su1' with a *substance* WORD error density of 27.083 is not depicted for reasons of space and layout.

Figure 4: Box plot of all *substance/scope* error densities.

distributions on the right-hand side are longer than on the left-hand side, meaning that the distributions are right-skewed. Most error densities stretch over a wide range of values. Every performance contains errors with *substance* WORD.

4.2 Correlation

To answer RQ2 (*What is the relationship between errors described using the Scope – Substance error taxonomy and the human ratings in the four dimensions of Grammar, Vocabulary, Coherence & Cohesion, and Task Achievement?*), the error densities were put into relation to the total fair ratings and to the fair ratings on each of the four dimensions. The relationship was analysed by means of correlation first, and in a following step, by means of regression analysis.

	TOT.Fair		GR.Fair		VOC.Fair		COH.Fair		TA.Fair	
	ρ	r	ρ	r	ρ	r	ρ	r	ρ	r
ED_TOT	$-.667^{**}$	$-.618^{**}$	$-.720^{**}$	$-.684^{**}$	$-.569^{**}$	$-.551^{**}$	$-.675^{**}$	$-.602^{**}$	$-.416^{**}$	$-.378^{**}$
ED_Su1	$-.571^{**}$	$-.561^{**}$	$-.623^{**}$	$-.610^{**}$	$-.513^{**}$	$-.532^{**}$	$-.522^{**}$	$-.501^{**}$	$-.369^{**}$	$-.370^{**}$
ED_Su2	$-.116$	$-.081$	$-.204$	$-.187$	$-.050$	$.003$	$-.252^{+}$	$-.184$	$.031$	$.055$
ED_Su3	$-.572^{**}$	$-.527^{**}$	$-.561^{**}$	$-.535^{**}$	$-.453^{**}$	$-.444^{**}$	$-.549^{**}$	$-.540^{**}$	$-.418^{**}$	$-.349^{*}$
ED_Su4	$-.123$	$-.097$	$-.118$	$-.052$	$-.024$	$.004$	$-.168$	$-.141$	$-.197$	$-.144$
ED_Su5	$-.535^{**}$	$-.569^{**}$	$-.471^{**}$	$-.530^{**}$	$-.455^{**}$	$-.498^{**}$	$-.438^{**}$	$-.443^{**}$	$-.554^{**}$	$-.552^{**}$
ED_Su9	$.134$	$.096$	$.053$	$.044$	$.091$	$.070$	$.084$	$.014$	$.238$	$.201$
ED_Sc1	$-.020$	$-.104$	$-.076$	$-.180$	$-.147$	$-.190$	$-.013$	$-.079$	$.089$	$.073$
ED_Sc2	$-.284^{+}$	$-.317^{*}$	$-.280^{+}$	$-.314^{*}$	$-.314^{*}$	$-.325^{*}$	$-.364^{*}$	$-.390^{**}$	$-.157$	$-.105$
ED_Sc3	$-.582^{**}$	$-.528^{**}$	$-.604^{**}$	$-.583^{**}$	$-.517^{**}$	$-.501^{**}$	$-.546^{**}$	$-.506^{**}$	$-.364^{*}$	$-.293^{*}$
ED_Sc4	$-.003$	$.015$	$-.105$	$-.082$	$-.060$	$-.008$	$-.111$	$-.025$	$.135$	$.112$
ED_Sc5	$-.520^{**}$	$-.445^{**}$	$-.465^{**}$	$-.423^{**}$	$-.304^{*}$	$-.275^{+}$	$-.496^{**}$	$-.395^{**}$	$-.532^{**}$	$-.481^{**}$

$N = 48$. Codes for significance levels (2-tailed): $^{+}\,p < .1$. $^{*}\,p < .05$. $^{**}\,p < .01$.
ρ = Spearman's rank-order correlation coefficient, r = Pearson's product-moment correlation coefficient

Table 4: Correlation between error densities and Fair Measures.

Table 4 shows the results of the correlation analysis. The following descriptions refer to Spearman's ρ.

Total error density

The total error density (ED_TOT) has highly significant correlations with all Fair Measure variables. The relationships are inverse. Performances with a higher overall fair rating tend to have a lower total error density.

***Substance* error densities**

The error densities for *substance* WORD (ED_Su1), CLAUSE (ED_Su3), and TEXT (ED_Su5) have significant, moderate to strong relationships with all Fair Measure variables. The *substance* PHRASE (ED_Su2), SENTENCE (ED_Su4), and PUNCTUATION error densities (ED_Su9) are not significantly correlated with any of the Fair Measures.

***Scope* error densities**

The error densities for *scope* CLAUSE (ED_Sc3) and *scope* TEXT (ED_Sc5) have moderate to strong correlations with all Fair Measure variables. The *scope* PHRASE error density (ED_Sc2) has weak correlations with the dimensions Vocabulary and Cohesion & Coherence, has no (significant) connection with the Task Achievement Fair Measure, and only shows a negative trend ($p < 0.1$) for Total and Grammar Fair Measure. ED_Sc3 has the predominant role for all Fair Measure variables apart from Task Achievement. ED_Sc5 has the strongest association with this variable. The *scope* WORD (ED_Sc1) and SENTENCE error densities (ED_Sc4) are not significantly correlated with any of the Fair Measures.

4.3 Regression

In contrast to correlation, which only expresses the strength of the relationship between two variables, regression provides additional information about the change in the dependent variable when the values in the independent variable(s) change. For this reason, we are using the regression coefficients to express sensitivity of the rating system to individual aspects of errors.

outcome	**TOT.Fair**		**GR.Fair**		**VOC.Fair**		**COH.Fair**		**TA.Fair**	
model	**1**		**2**		**3**		**4**		**5**	
intercept	6.087**	(.394)	6.338**	(.400)	5.925**	(.448)	5.980**	(.454)	5.828**	(.485)
ED_TOT	**−.121****	(.021)	**−.146****	(.022)	**−.111****	(.024)	**−.129****	(.025)	**−.090****	(.026)
R^2	.423		.502		.317		.379		.199	
adj. R^2	.411		.491		.302		.366		.181	
robRSE	1.007		1.036		1.149		1.178		1.280	
model	**6**		**7**		**8**		**9**		**10**	
intercept	5.350**	(.353)	5.868**	(.443)	5.213**	(.476)	5.514**	(.460)	5.121**	(.398)
ED_Su1	**−.091****	(.024)	**−.119****	(.030)	**−.091****	(.032)	**−.099****	(.031)	−.050^{+}	(.027)
ED_Su2	−.036	(.085)	−.126	(.106)	.074	(.114)	−.097	(.110)	.089	(.095)
ED_Su3	**−.314****	(.098)	**−.289***	(.121)	−.247^{+}	(.130)	**−.342***	(.128)	−.101	(.114)
ED_Su4	−.015	(.241)	.149	(.300)	.210	(.323)	−.175	(.313)	−.173	(.270)
ED_Su5	**−.445****	(.156)	**−.419***	(.196)	**−.450***	(.209)	−.260	(.202)	**−.889****	(.175)
ED_Su9	.140^{+}	(.071)	.063	(.087)	.095	(.094)	.123	(.094)	**.195***	(.080)
R^2	.660		.600		.479		.523		.600	
adj. R^2	.610		.541		.403		.453		.541	
robRSE	.796		.983		1.062		1.069		.908	
model	**11**		**12**		**13**		**14**		**15**	
intercept	5.808**	(.393)	6.088**	(.427)	5.808**	(.497)	5.835**	(.476)	5.318**	(.545)
ED_Sc1	−.014	(.071)	−.057	(.077)	−.061	(.090)	.005	(.088)	.055	(.099)
ED_Sc2	−.145	(.083)	−.159^{+}	(.089)	−.185^{+}	(.105)	**−.250***	(.101)	−.052	(.115)
ED_Sc3	**−.162****	(.040)	**−.182****	(.043)	**−.144****	(.050)	**−.151****	(.049)	**−.115***	(.055)
ED_Sc4	.052	(.078)	−.003	(.085)	.025	(.098)	.024	(.094)	.104	(.108)
ED_Sc5	**−.163****	(.041)	**−.178****	(.044)	**−.105***	(.051)	**−.168****	(.050)	**−.176****	(.056)
R^2	.527		.554		.360		.467		.310	
adj. R^2	.471		.501		.284		.414		.228	
robRSE	.958		1.029		1.181		1.137		1.285	

$N = 48$. Codes for significance levels (2-tailed): $^{+}\,p < .1$. $^{*}\,p < .05$. $^{**}\,p < .01$.

Models in the second column (1, 6, 11) have the outcome variable TOT.Fair, models in the third column (2, 7, 12) GR.Fair, and so on. robRSE = robust residual standard error. Parameter estimation method: MM-estimator. The coefficients are unstandardized, the standard errors are printed in parentheses.

Table 5: Regression models for error densities predicting Fair Measures.

Total error density

The simple linear regression models for the total error density (ED_TOT) predicting the Fair Measures are given in models 1 to 5 in Table 5. The relationships are negative, meaning that a higher error density is associated with lower ratings in all dimensions. The total error density has the strongest impact on the Grammar Fair Measure and the weakest on the Task Achievement Fair Measure. The (estimated) regression function from model 2 can be formulated as follows:

$$\text{Grammar Fair Measure: GR.Fair} = 6.338 - 0.146 \cdot \text{ED\_TOT}$$

Using descriptive language, this equation can be described as follows: Comparing two learners of English at grade 8 that differ by 1 in the total error density, the Grammar Fair Measure is expected to differ by 0.146, and the learner with the lower error density has the higher rating. Theoretically, a performance without any errors would have a rating of 6.338.

All models have limited accuracy. The proportion of the variance in the outcome variable Task Achievement Fair Measure that is predictable by the total error density is even as low as 20% ($R^2 = 0.199$). The highest coefficient of determination can be found in the regression model on Grammar Fair Measure ($R^2 = 0.502$)

Substance error densities

The multiple linear regression models between the *substance* error densities (ED_SuX) and the Fair Measures are shown in models 6 to 10 in Table 5. The error densities for *substance* WORD (ED_Su1), *substance* CLAUSE (ED_Su3), and TEXT (ED_Su5) are significant regressors in almost all models. ED_Su5 has the largest coefficients. Its impact on the Task Achievement dimension is especially strong. A difference of 1 in the error density is associated with a difference in 0.889 in the Task Achievement rating. In contrast, the error density for *substance* PUNCTUATION (ED_Su9) has a significant positive coefficient in the Task Achievement model (model 10). This is the only significant positive coefficient.

Similar to the regression with the total error density, all models have limited accuracy. The highest proportion of the variance in the dependent variables that is explained by the *substance* error densities is 66% (model 6). This time, the model with Task Achievement has a relatively high R^2 (0.6).

Scope error densities

The regression models for the *scope* error densities (ED_ScY) predicting the Fair Measures are shown in models 11 to 15 in Table 5. The error densities for *scope* CLAUSE

(ED_Sc3) and TEXT (ED_Sc5) have a significant negative impact on all Fair Measure variables. The situation is less clear for the *scope* PHRASE (ED_Sc2) error density. While the coefficients have a relatively large, negative value, the variable is only significant at the 0.05 level (two-tailed) for the Coherence & Cohesion Fair Measure (model 14). However, there is a similar trend ($p < 0.1$) on the other ratings (except for Task Achievement). For none of the five models are the *scope* WORD and SENTENCE error densities statistically distinguishable from zero.

The accuracy of the models is again not high. The highest coefficient of determination is 0.554 in the regression on Grammar Fair Measure (model 12), and the lowest is 0.31 for Task Achievement Fair Measure (model 15).

5 Discussion

5.1 Areas of Difficulty for Learners

The average total error density is approximately 17, meaning that, on average, each performance contains around 17 errors per 100 words. This leads to the conclusion that writing, especially producing accurate language, is a great difficulty for Austrian learners at grade 8.

The majority of the errors concern the unit WORD, which is not surprising considering that most of the errors relate to prepositions, spelling, personal pronouns, lexical choice and capitalization (cf. Pibal 2012), most of which are correctable by changing a single word. The high frequency of punctuation errors, most of which concern commas, may have two causes. First, the errors may result from L1 interference (language transfer). For example, in German it is required to surround every subordinate clause with commas, which is not the case for English. Additionally, the incorrect addition or omission of a comma usually does not lead to a change in meaning or to a communication problem and, thus, it is reasonable to conclude that students do not pay a lot of attention to (certain) comma rules and the related errors. For the same reasons, teachers might not consider it necessary to give detailed feedback on such errors or provide clear instructions on the correct usage.

Most errors involve *scope* beyond PHRASE (scope CLAUSE or SENTENCE constitute 51% of all non-norm adequate forms) and errors with *scope* TEXT turn out to be the second most common category (21%). This leads to the conclusion that a lot of pupils are not yet able to see their text as a meaningful circle of ideas that are interconnected. Thus, recognising the repercussions an error has on textual context seems to be a major challenge for the learners and should be focused on in teaching.

5.2 Error Densities and Fair Measures: Implications for Validity

It turns out that a high error density is typical of performances with a low overall rating, thus indicating that errors and writing competence are connected. This finding is not unexpected and has been observed in numerous studies (e.g., Bıdık 2016; Frey and Heringer 2007; Homburg 1984; Weltig 2004; Wolfe-Quintero, Inagaki, and Kim 1998).

In what follows, only the results of the regression analyses are discussed because they provide more information than correlation. While correlation, which only expresses the strength of the relationship between two variables, regression also indicates the change in the dependent variable brought about by a change of the values in the independent variable(s). For this reason, we use the regression coefficients to express sensitivity of the rating system to individual errors.

A meaningful observation concerns the intercept of the regression models. The intercept is the expected mean value of the dependent variable, here the Fair Measure, when all independent variables, here error densities, are zero. In this context, it can be interpreted as the rating of a performance that does not contain any errors. For none of the models is the intercept 7, which would be the perfect rating. It appears that errors alone do not fully explain the rating, which is also evident by the limited regression accuracy.

The total error density has a relatively strong, significant impact on most of the dependent variables. For example, a difference in the value of the total error density of 8 is connected to a difference in one level of the 7-level Total Fair Measure (see appendix). Given that 50% of the performances have more than 16.36 errors per 100 words, this is a considerable impact.

However, the total error density has only a small impact on the rating in the Task Achievement dimension. Task achievement refers to the ability to produce texts that respond to the tasks in a clear and meaningful way and to elaborate and expand ideas meaningfully (Kulmhofer and Siller 2018). Consequently, the total number of errors identified using the *Scope – Substance* error taxonomy should not have a strong impact on this dimension. Since the regression weight is very small (ß = -0.090, p < 0.01, regression model 5) and the accuracy very low (R2 = 0.199), raters seem to consider different things when making their judgments, such as whether all content points from the prompt have been mentioned or whether they have been elaborated or not (cf. appendix and the extended scale in Gassner, Mewald, and Sigott (2008, 25)).

Among the error features which exert a strong statistical influence on the ratings are the *substance* WORD (ED_Su1), CLAUSE (ED_Su3) and TEXT (ED_Su5) error densities. *substance* WORD affects ratings on three of the four rating dimensions, more strongly on Grammar than on Vocabulary. *substance* CLAUSE also affects ratings on Grammar and, more strongly, on Cohesion and Coherence. Not surprisingly, *substance* TEXT has the strongest statistical effect on Task Achievement but also strongly affects the ratings on Grammar and Vocabulary. The effect of *substance* TEXT on Task Achievement is in line with expectations based on the fact that *substance* TEXT errors are the only error type that directly captures problems with task achievement.

Contrary to the expectation that all error types have a negative effect on the rating, the *substance* PUNCTUATION error density has a positive, statistically significant, regression weight for the Task Achievement dimension (ß $= 0.195$, $p < 0.05$, regression model 10). The incidence of punctuation errors, then, increases with text quality. This is presumably due to the fact that the higher the quality of the text, the more complex the structures, but the more punctuation is needed. Conceivably, this increases the writers' chances of making punctuation errors. In contrast to the other error types, punctuation errors, while frequent, do not affect the ratings negatively. A similar phenomenon was observed by Bıdık (2016).

Concerning the amount of context needed to detect an error, i.e., *scope*, the error categories CLAUSE (ED_Sc3) and TEXT (ED_Sc5) are the ones with a significant impact on the ratings. *Scope* CLAUSE affects all the ratings, but most strongly the Grammar rating, while *scope* TEXT also affects all the ratings but most strongly the Grammar and Task Achievement ratings.

These results constitute detailed information about aspects of student writing that seem to be important in determining quality assessments of the writing. The *substance* categories which are associated with the ratings provide information on what kind of changes in the performances would lead to higher ratings. The *scope* categories could constitute a basis for formulating feedback to students by pointing out to them how much of their text they need to consider in order to avoid particular *substance* errors. Thus, the results could also be seen as a contribution to formative assessment.

In all, the results make a contribution to identifying the construct of the Austrian E8 Standards Writing test. They address context validity by identifying features that the raters seem to be sensitive to, and thus constitute sources of task difficulty. At the same time, the results also speak to theory-based validity as they identify errors that Austrian adolescent writers need to avoid, i.e., areas of competence that they need to develop, in order to enhance their writing.

However, it has to be borne in mind that the error categories studied do not explain the entire variance in the ratings. This indicates that factors other than errors, namely positive features, do seem to play a role in the assessment as well. Identifying such positive features is a worthwhile and necessary task for future research.

6 Conclusion

This study analysed errors by means of the *Scope – Substance* error taxonomy in a sample of 50 written performances from the 2009 baseline study for the E8 Standards Writing Test. The primary aim was to contribute to the validation of this large-scale assessment by studying the relationship between errors (described using the *Scope – Substance* error taxonomy) and human ratings awarded to writing performances (RQ2). An additional aim was to examine the data to determine the most common errors in the writing of 14-year-old Austrian learners of English (RQ1).

The study provides insight into the areas in which pupils have difficulties. In general, writing, especially producing accurate language, poses a great difficulty for Austrian learners of English at grade 8. Most of the problems concern the unit WORD (59%), followed by PUNCTUATION (14%). Regarding *scope*, learners have problems considering the constraints of wider context (clause, sentence, text) to produce accurate language.

These results constitute baseline information about error occurrence in adolescent L2 English writing of L1 German-speaking learners. Unlike in the US, where data on error occurrence has been monitored over decades (A. A. Lunsford and K. J. Lunsford 2008; K. C. Wilcox, Yagelski, and F. Yu 2013), such information has not been collected for L2 English writing in the German-speaking context. Now that this study has established the baseline, it would be interesting to follow the development of error incidence in L2 English writing in the German-speaking area in the future.

The data indicate that errors and assessments of writing competence in the E8 Standards Writing Test are connected. In line with prior studies, this study also showed that as writers become more proficient, they tend to produce increasingly accurate language. In this study, this manifests itself as a negative relationship between human ratings and the presence of errors identified by means of the *Scope – Substance* error taxonomy. A low error density is associated with high ratings and a high error density with low ratings. *substance* WORD, CLAUSE, and TEXT error densities play an important role in the rating in most dimensions; errors with a larger *scope* also have a strong effect. By highlighting aspects of errors to which raters seem to be sensitive, these findings constitute evidence of context validity. At the same time, the findings are relevant to theory-based validity by concretising areas of competence that learners need to develop in order to receive higher ratings. While errors are important determinants of the ratings, additional factors must be at play as the accuracy of the regression models is not perfect. It seems that raters do take errors into consideration in the assessment, but other variables, presumably positive features, also contribute to their decisions. This should in fact be the case since the rating scale refers to negative as well as positive features.

The study has implications for practice. On the one hand, it provides a basis for refining the descriptors of the E8 rating scale for writing by adding the aspects of error to which raters seem to be particularly sensitive. This could help to make the rating easier and contribute to reliability and validity. Making these aspects of errors explicit in rater training should also prove helpful in efforts to maximise reliability and validity. On the other hand, if aspects of errors to which raters are sensitive are communicated to teachers, the competences required to avoid such errors can be focused on in the teaching of writing. This can facilitate positive washback, which is an important aim of Austrian Educational Standards testing.

References

Bıdık, Buket. 2016. "Refining the SD Error Taxonomy: A Descriptive Analysis of Grammatical Errors in Samples by Turkish EFL Learners." MA thesis, Department of English and American Studies, University of Klagenfurt.

Brants, Thorsten. 2000. "Inter-Annotator Agreement for a German Newspaper Corpus." In *Second International Conference on Language Resources and Evaluation LREC-2000*. Athens, Greece. http://www.coli.uni-saarland.de/~thorsten/publications/Brants-LREC00.pdf.

Cizek, Gregory, and Michael Bunch. 2007. *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, California: SAGE Publications, Inc.

Cook, Vivian J. 1993. *Linguistics and Second Language Acquisition*. Basingstoke: Macmillan.

Corder, Pit. 1974. "Error Analysis." In *Techniques in Applied Linguistics*, edited by John P. B. Allen and Pit Corder, 122–154. The Edinburgh Course in Applied Linguistics 3. London: Oxford University Press.

Díaz-Negrillo, Ana, and Jesús Fernández-Domínguez. 2006. "Error Tagging Systems for Learner Corpora." *RESLA* 19: 83–102.

Dobrić, Nikola. 2015. "Quality Measurements of Error Annotation - Ensuring Validity Through Reliability." *The European English Messenger* 24 (1): 36–42.

Dobrić, Nikola. 2020. "Rater Cognition and Errors: A Corpus-Based Approach to Validating Writing Assessment." Habilitationsschrift, University of Klagenfurt.

Dobrić, Nikola, and Günther Sigott. 2014. "Towards an Error Taxonomy for Student Writing." *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 19 (2): 111–118. http://tujournals.ulb.tu-darmstadt.de/index.php/zif/article/download/35/32.

Dobrić, Nikola, Günther Sigott, Gašper Ilc, Vesna Lazović, Hermann Cesnik, and Andrej Stopar. 2021. "Errors as Indicators of Writing Task Difficulty at the Slovene General Matura in English." *International Journal of Applied Linguistics*, 1–17. https://doi.org/10.1111/ijal.12345.

Ellis, Rod, and Gary Barkhuizen. 2005. *Analysing Learner Language*. Oxford applied linguistics. Oxford: Oxford University Press.

Fliri, Benjamin, ed. 2018. *ÖGSD Tagungsberichte Vol. 4*: 10. Nachwuchstagung. Sprachendidaktik: Der Wissenschaftliche Nachwuchs Im Dialog. (Proceedings of the 10th ÖGSD Young

Researchers' Conference). Graz: ÖGSD. https://www.oegsd.at/wp-content/uploads/2020/08/2018-Nachwuchstagung-Bericht-und-extended-Abstracts.pdf.

Frey, Evelyn, and Hans Jürgen Heringer. 2007. "Automatische Bewertung Schriftlicher Lernerproduktionen." *Linguistische Berichte* 211: 331–345.

Gassner, Otmar, C. Mewald, R. Brock, F. Lackenbauer, and Klaus Siller. 2011. *Testing Writing for the E8 Standards: Technical Report 2011*. Salzburg: BIFIE Salzburg. Accessed February 08, 2018. http://www.bifie.at/wp-content/uploads/2017/05/bist_Technical-Report2-E8_2011-09-26.pdf.

Gassner, Otmar, C. Mewald, and Günther Sigott. 2008. "Testing Writing: Specifications for the E8-Standards Writing Tests." LTC Technical Report 4. Accessed March 07, 2020. https://www.aau.at/wp-content/uploads/2018/03/LTC_Technical_Report_4.pdf.

Granger, Syviane. 2003. "Error-Tagged Learner Corpora and CALL: A Promising Synergy." *CALICO Journal* 20 (3): 465–480.

Hafner, Samuel. 2018a. "Analyzing Errors in Written Performances by Means of the Scope – Substance Error Taxonomy and Investigating Their Influence on Human Ratings." Diploma thesis, Department of English, University of Klagenfurt. https://netlibrary.aau.at/urn:nbn:at:at-ubk:1-36593

Hafner, Samuel. 2018b. "Refining the Scope – Substance Error Taxonomy by Means of an Agreement Study." In *ÖGSD Tagungsberichte Vol. 4: 10. Nachwuchstagung. Sprachendidaktik: Der Wissenschaftliche Nachwuchs Im Dialog*, edited by Benjamin Fliri et al., 18–20. Graz: ÖGSD.

Hammarberg, Björn. 1974. "The Insufficiency of Error Analysis." *International Review of Applied Linguistics in Language Teaching* 12: 185–192.

Homburg, Taco Justus. 1984. "Holistic Evaluation of ESL Compositions: Can It Be Validated Objectively?" *TESOL Quarterly* 18 (1): 87–107. https://doi.org/10.2307/3586337.

IQS. n.d. "Ausgangsmessung, 8. Schulstufe (2009)." Accessed May 18, 2021. https://www.iqs.gv.at/themen/bildungsforschung/forschungsdatenbibliothek/daten-der-bildungsstandardueberpruefungen/fdb-ausgangsmessung-8-schulstufe-2009.

James, Carl. 1998. *Errors in Language Learning and Use: Exploring Error Analysis*. Applied Linguistics and Language Study. London, New York: Longman.

Khansir, Ali Akbar. 2012. "Error Analysis and Second Language Acquisition." *Theory and Practice in Language Studies* 2 (5): 1027–1032. https://doi.org/10.4304/tpls.2.5.1027-1032.

Koller, Manuel, and Werner A. Stahel. 2011. "Sharpening Wald-Type Inference in Robust Regression for Small Samples." *Computational Statistics & Data Analysis* 55 (8): 2504–2515. https://doi.org/10.1016/j.csda.2011.02.014.

Koller, Manuel, and Werner A. Stahel. 2017. "Nonsingular Subsampling for Regression S Estimators with Categorical Predictors." *Computational Statistics* 32 (2): 631–646. https://doi.org/10.1007/s00180-016-0679-x.

Kulmhofer, Andrea, and Klaus Siller. 2018. "The Development of the Austrian Educational Standards Test for English Writing at Grade 8." In *Language Testing in Austria: Taking Stock/Sprachtesten in Österreich: Eine Bestandsaufnahme*, edited by Günther Sigott, 129–145. Frankfurt am Main: Peter Lang.

Lennon, Paul. 1991. "Error: Some Problems of Definition, Identification, and Distinction." *Applied Linguistics* 12 (2): 180–196. https://doi.org/10.1093/applin/12.2.180.

Lennon, Paul. 2008. "Contrastive Analysis, Error Analysis, Interlanguage." In *Bielefeld Introduction to Applied Linguistics*, edited by S. Gramley and V. Gramley, 51–60. Bielefeld: Aisthesis.

Lumley, Tom. 2005. *Assessing Second Language Writing: The Rater's Perspective*. Berlin: Peter Lang.

Lunsford, Andrea A., and Karen J. Lunsford. 2008. "'Mistakes Are a Fact of Life': A National Comparative Study." *College Composition and Communication* 59 (4): 781–806. https://www.jstor.org/stable/20457033?seq=1#metadata_info_tab_contents.

Maechler, Martin, Peter Rousseeuw, Croux, Christophe, Todorov, Valentin, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Eduardo L. T. Conceicao, and Maria Anna Di Palma. 2021. *Robustbase: Basic Robust Statistics R Package Version 0.93-7*. https://CRAN.R-project.org/package=robustbase.

Pibal, Florian. 2012. "Identifying Errors in the Written Manifestations of Austrian English Learner Language at 8th-Grade Secondary Level and Their Influence on Human Ratings." MA thesis, Department of English and American Studies, University of Klagenfurt.

Pibal, Florian, Günther Sigott, and Hermann Cesnik. 2018. "The Role of Error in Assessing Writing in the National Educational Standards Baseline Test." In *Language Testing in Austria: Taking Stock/Sprachtesten in Österreich: Eine Bestandsaufnahme*, edited by Günther Sigott, 419–443. Frankfurt am Main: Peter Lang.

Plaban, Bhowmick, Mitra Pabitra, and Basu Anupam. 2008. "An Agreement Measure for Determining Inter-Annotator Reliability of Human Judgements on Affective Text." In *Proceedings of the Workshop on Human Judgements in Computational Linguistics*, edited by Ron Artstein, Gemma Boleda, Frank Keller, and Sabine Schulte im Walde, 58–65. Manchester, UK: Coling 2008 Organizing Committee. http://aclweb.org/anthology/W08-12.

Potter, W. James, and Deborah Levine-Donnerstein. 1999. "Rethinking Validity and Reliability in Content Analysis." *Journal of Applied Communication Research* 27 (3): 258–284. https://doi.org/10.1080/00909889909365539.

Quirk, Randolph, Sidney Greenbaun, Geoffrey Leech, and Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Saville-Troike, Muriel. 2006. *Introducing Second Language Acquisition*. Cambridge introductions to language and linguistics. Cambridge: Cambridge University Press.

Schachter, Jacquelyn. 1974. "An Error in Error Analysis." *Language Learning* 24 (2): 205–214. https://doi.org/10.1111/j.1467-1770.1974.tb00502.x.

Shmueli, Galit. 2010. "To Explain or to Predict?" *Statistical Science* 25 (3): 289–310. https://doi.org/10.1214/10-STS330.

Sigott, Günther. 2004. *Towards Identifying the C-Test Construct*. Frankfurt am Main: Peter Lang.

Sigott, Günther, ed. 2018. *Language Testing in Austria: Taking Stock/Sprachtesten in Österreich: Eine Bestandsaufnahme*. Frankfurt am Main: Peter Lang.

Sigott, Günther, Hermann Cesnik, and Nikola Dobrić. 2016. "Refining the Scope – Substance Error Taxonomy: A Closer Look at Substance." In *Corpora in Applied Linguistics: Current Ap-*

proaches, edited by Nikola Dobrić, Eva-Maria Graf, and Alexander Onysko, 79–94. Newcastle upon Tyne, UK: Cambridge Scholars Publishing.

Steinkellner, Florian. 2018. "Towards a Refined Version of the Scope – Substance Error Taxonomy." In *ÖGSD Tagungsberichte Vol. 4: 10. Nachwuchstagung. Sprachendidaktik: Der Wissenschaftliche Nachwuchs Im Dialog*, edited by Benjamin Fliri et al., 46–48. Graz: ÖGSD.

Weir, Cyril J. 2005. *Language Testing and Validation: An Evidence-Based Approach*. Research and practice in applied linguistics. Basingstoke: Palgrave Macmillan.

Weltig, Matthew S. 2004. "Effects of Language Errors and Importance Attributed to Language on Language and Rhetorical-Level Essay Scoring." *Spaan Fellow Working Papers in Second or Foreign Language Assessment* 2: 53–81.

Wilcox, Kristen Campbell, Robert Yagelski, and Fang Yu. 2013. "The Nature of Error in Adolescent Student Writing." *Read Writ* 27 (6): 1073–1094. https://doi.org/10.1007/s11145-013-9492-x.

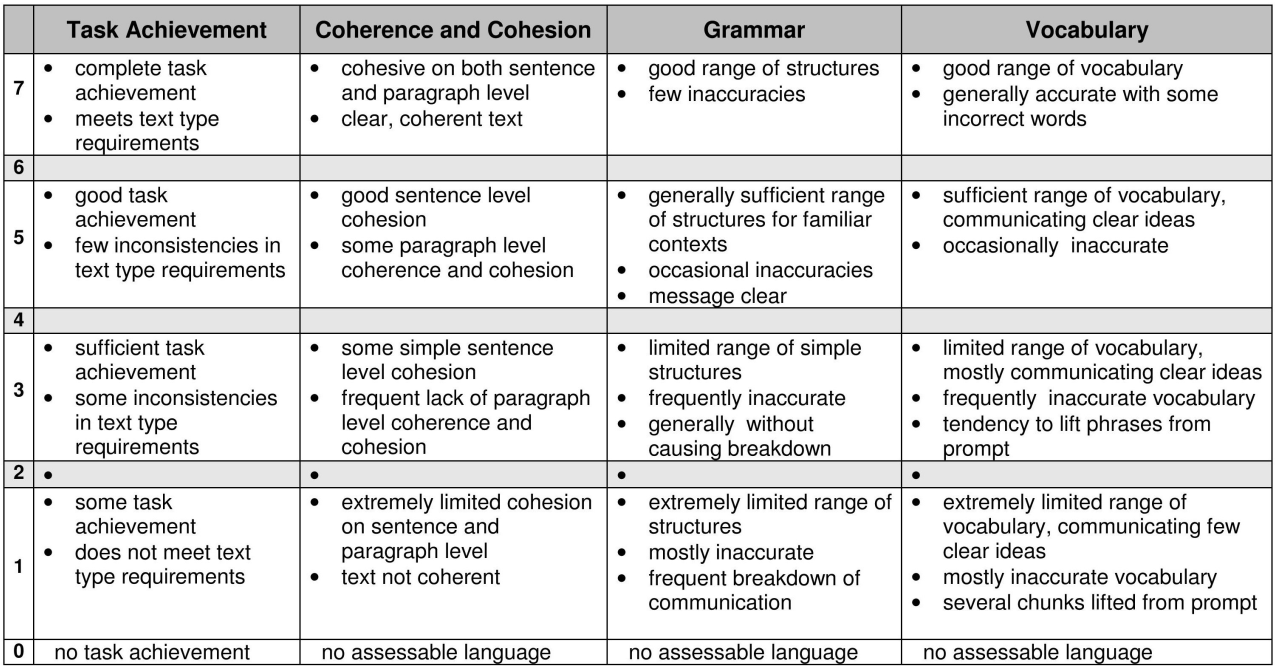
Wilcox, Rand R. 2012. *Introduction to Robust Estimation and Hypothesis Testing*. 3rd ed. Statistical Modeling and Decision Science. Amsterdam: Elsevier.

Wolfe-Quintero, Kate, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy, & Complexity*. Technical report / Second Language Teaching & Curriculum Center 17. Honolulu, Hawaii: University of Hawai'i Press.

Yu, Xiu. 2017. "On the Avoidance Phenomenon in Writing." *Journal of Language Teaching and Research* 8 (5): 948–952. https://doi.org/10.17507/jltr.0805.15.

Appendix

E8 Writing Rating Scale (version May 2008)



	Task Achievement	Coherence and Cohesion	Grammar	Vocabulary
7	• complete task achievement • meets text type requirements	• cohesive on both sentence and paragraph level • clear, coherent text	• good range of structures • few inaccuracies	• good range of vocabulary • generally accurate with some incorrect words
6				
5	• good task achievement • few inconsistencies in text type requirements	• good sentence level cohesion • some paragraph level coherence and cohesion	• generally sufficient range of structures for familiar contexts • occasional inaccuracies • message clear	• sufficient range of vocabulary, communicating clear ideas • occasionally inaccurate
4				
3	• sufficient task achievement • some inconsistencies in text type requirements	• some simple sentence level cohesion • frequent lack of paragraph level coherence and cohesion	• limited range of simple structures • frequently inaccurate • generally without causing breakdown	• limited range of vocabulary, mostly communicating clear ideas • frequently inaccurate vocabulary • tendency to lift phrases from prompt
2	•	•	•	•
1	• some task achievement • does not meet text type requirements	• extremely limited cohesion on sentence and paragraph level • text not coherent	• extremely limited range of structures • mostly inaccurate • frequent breakdown of communication	• extremely limited range of vocabulary, communicating few clear ideas • mostly inaccurate vocabulary • several chunks lifted from prompt
0	no task achievement	no assessable language	no assessable language	no assessable language

Figure 5: E8 Writing Rating Scale (version May 2008) (Gassner, Mewald, and Sigott 2008, 24)