Automated Scoring of a German Written Elicited Imitation Test

Mihail Chifligarov[1], Jammila Laâguidi[1], Max Schellenberg[1], Alexander Dill[1], Anna Timukova[2,4], Anastasia Drackert[3,4] and **Ronja Laarmann-Quante[1]**
Ruhr University Bochum, Germany
firstname.lastname@rub.de
[1]Faculty of Philology, Department of Linguistics
[2]University Language Centre (ZFA)
[3]Faculty of Philology, Institute for German Language and Literature
[4]g.a.s.t. (Society for Academic Study Preparation and Test Development)

Abstract

We present an approach to the automated scoring of a German Written Elicited Imitation Test, designed to assess literacy-dependent procedural knowledge in German as a foreign language. In this test, sentences are briefly displayed on a screen and, after a short pause, test-takers are asked to reproduce the sentence in writing as accurately as possible. Responses are rated on a 5-point ordinal scale, with grammatical errors typically penalized more heavily than lexical deviations. We compare a rule-based model that implements the categories of the scoring rubric through hand-crafted rules, and a deep learning model trained on pairs of stimulus sentences and written responses. Both models achieve promising performance with quadratically weighted kappa (QWK) values around .87. However, their strengths differ – the rule-based model performs better on previously unseen stimulus sentences and at the extremes of the rating scale, while the deep learning model shows advantages in scoring mid-range responses, for which explicit rules are harder to define.

1 Introduction

The Written Elicited Imitation Test (WEIT) is a computer-based test designed to measure procedural linguistic knowledge in writing. In this test, learners briefly view sentences in the target language and, after a short pause, reproduce them from memory by typing. Responses are then rated on an ordinal scale based on how closely they resemble the original sentences.

Like any assessment that relies on scoring by human raters, the WEIT can benefit greatly from automation. An automated scoring system would significantly improve efficiency by enabling the rapid evaluation of large numbers of responses without the time and effort required by human raters. This would, in turn, allow for immediate feedback, an advantage in both instructional and research contexts. Automation also ensures greater consistency and objectivity by applying scoring criteria uniformly and eliminating potential rater bias. In addition, automated systems can provide fine-grained data on error patterns and processing behavior, offering deeper insight into learners' procedural language skills.

In this paper, we investigate the automated scoring of a German WEIT. The responses in our dataset are scored using a rubric that assigns a score between 0 and 4, based on deviations in spelling, grammar, and vocabulary (see Section 3.2). There are two main approaches to automating this process: a rule-based approach, in which categories from the scoring rubric are implemented explicitly, and a deep learning approach, in which a model learns implicitly which scores to apply based on training pairs of stimulus and response sentences.

In educational settings, transparency and explainability are important considerations. From this perspective, rule-based models are preferable as they allow for a clearer justification of scoring decisions and can offer more detailed feedback to learners by pinpointing specific types of deviations. However, rule-based systems can be limited in flexibility, particularly when dealing with edge cases or language exceptions. In contrast, deep learning models may be better suited to capturing subtle patterns in learner responses (e.g. to what extent a word substitution affects the overall meaning of the sentence), but often lack transparency and may struggle to generalize to previously unseen stimulus sentences.

This paper presents a rule-based scoring model for the WEIT, built on general principles derived from the scoring rubric, and compares it to a deep learning model trained on stimulus-response pairs. We hypothesize that (a) the deep learning model will outperform the rule-based model on cases where differences between descriptors in the scor-

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 237–247

ing rubric are rather subtle and hard to capture as explicit rules, and (b) the rule-based scoring model will generalize better to new, previously unseen stimulus sentences.

The main contributions of this paper are twofold. First, we explore the feasibility of automating the scoring of a German WEIT using a detailed ordinal rating scale. Second, we provide a concrete case study for comparing the strengths and limitations of deep learning and rule-based methods in an educational assessment context.

Our code and data are available at: https://gitlab.ruhr-uni-bochum.de/vamos-cl/german-weit-automated-scoring.

2 Background and Related Work

In the following, we provide background on the use of elicited imitation tests and summarize previous research on their automated scoring.

2.1 Elicited Imitation Tests (EIT): Construct and Use

Elicited imitation tests (EIT) have been widely used and researched in the field of Second Language Acquisition as measures of two key constructs: global proficiency in a second or foreign language (Drackert, 2016; Kostromitina and Plonsky, 2022) and implicit language learning (Nikouee and Ranta, 2023). EITs exist in many languages and have been primarily employed in the oral mode (oral EIT, or OEIT) in which language learners listen to a number of sentences and then orally repeat them as accurately as possible after a short pause.

Recently, written elicited imitation tests (WEIT) have started to gain attention in language testing as a research tool (e.g. Sun et al., 2025) or as a measure of literacy-dependent procedural language knowledge (Timukova et al., submitted).

In the format that was used by Timukova et al. (submitted) in a large language testing research project, the sentences are briefly presented on a screen, and, after a pause, learners have to reproduce as much of the sentence as they can by typing their response into a text box. The pause is intended to reduce the influence of working memory and to promote active reconstruction of the stimulus rather than rote repetition. The construct of literacy-dependent procedural language knowledge measured by WEIT can be defined as automatized knowledge and skills required for the real-time reception and production of written language.

Inspired by and closely related to the well-established oral elicited imitation format (Ortega et al., 2002), the written test — despite being presented and completed in a different modality and incorporating a distinct scoring system to better capture the construct (see Section 3.2) — yields results of comparable difficulty and reliability.[1] However, it is considerably easier to develop, administer, and score, as no audio equipment is required at any stage. Scoring short written responses is also likely more practical and less time-consuming than scoring spoken responses when done by human raters.

2.2 Automated Scoring of EITs

While EITs, in principle, lend themselves well to automated scoring since the target response is known (i.e. exact repetition of the stimulus sentence), the difficulty of the automated scoring task largely depends on the scale or rubric used for rating responses that deviate from the target.

For the oral EIT, numerous studies have explored automated scoring of the test using automatic speech recognition (ASR), primarily employing a binary scale that codes whether the response matches the stimulus or not (e.g. Millard, 2011), or an interval scale, where, for example, one point is subtracted for each deviation in the response sentence (Graham et al., 2008; Lonsdale and Christensen, 2011). Once the learner utterances are accurately transcribed, automated scoring based on these scales is straightforward.

Besides binary and interval scales, ordinal scales exist where scores are determined qualitatively. In their meta-analysis, Yan et al. (2016) found that for the OEIT, ordinal rating scales were more effective at distinguishing speakers across proficiency levels than other scales. An established ordinal rating scheme for the OEIT is that of Ortega et al. (2002), where the score depends on how much of the stimulus sentence a learner was able to repeat:

- 0 points for minimal (one word), unintelligible responses or no repetition
- 1 point when half or less of the stimulus was repeated
- 2 points for changes to the original sentence in content or form that affected the meaning

[1] The tests used in the project showed difficulty indices of 0.41 (WEIT) and 0.49 (OEIT), and reliability coefficients (Cronbach's α) of .97 for both (N = 195).

- 3 points for accurate content repetition with some (un)grammatical changes
- 4 points for exact repetition with formal accuracy

Recent studies have investigated how automatically obtained scores based on objectively quantifiable features correlate with scores based on Ortega's ordinal scale. McGuire and Larson-Hall (2025) found high correlations with word error rate (WER), especially when looking at a participant's mean score across a whole test ($r = -0.969$). The correlation of WER and Ortega's scores across items, however, was lower ($r = -0.817$). Isbell et al. (2023) took further metrics such as Levenshtein distance into account and also mapped a combination of Percent Word Correct (PWC, exact matches) and Percent Meaning Correct (PMC, matching lemmata) to Ortega's 5-point scale (e.g. PWC $< 100\%$ and PMC $\geq 70\%$ = Score 3). They also found high correlations with Ortega's scores assigned by human raters (around $r = 0.9$ when aggregated across all items and around Spearman's $\rho = 0.8$ at the item level, depending on the metric and ASR service used).

In the present study, our aim is to implement an automated scoring procedure for a German WEIT, using an ordinal scale similar to that of Ortega et al. The scoring rubric will be presented in more detail in Section 3.2. It was specifically developed for the German WEIT, as no comparable schemes had yet existed. As the purpose of the WEIT is to test literacy-dependent procedural language knowledge, the rubric differs in some essential ways from that of Ortega et al. Our goal is to build a rule-based scoring model that implements the various categories from the rubric, rather than relying on purely quantitative measures such as WER. This scoring method is comparable to human raters' assessments in that it could provide learners with feedback about the scores they received based on the deviations in their responses. For comparison, we investigate how successful modern deep learning approaches are at approximating human ratings by implicitly learning to apply the scoring rubric.

3 Data

3.1 Data Collection

The data for our study was collected within a larger research project where the WEIT was used as a measure of literacy-dependent procedural knowledge (see Section 2.1). The 20 items included in the WEIT range from 6 to 16 words, or 8 to 24 syllables (see Appendix A for the full list of items). The test was completed by 195 university students who were learners of German (58.1% female, 41.9% male) between the ages of 18 and 40 ($M = 25.46$, $SD = 3.92$). The participants represented 47 different native languages, with Russian ($n = 30$), Turkish ($n = 23$), English and Spanish ($n = 14$ each) being the most frequent. Most participants self-assessed their language skills to be somewhere between A2 and C1.

3.2 Scoring Rubric

An ordinal scoring rubric for the German WEIT was developed for the purposes of the project. It follows the rubric of Ortega et al. (2002) in that responses are scored based on how closely they resemble the stimulus sentences. A key difference between the WEIT rubric and the OEIT rubric already addressed in Section 2.2 is the altered role of *meaning* and *grammar*. Since rule-governed morphological and syntactic sequences are central to the construct of procedural knowledge measured by the WEIT, grammatical deviations carry more weight. Hence, the rubric distinguishes between lexical and grammatical deviations from the original, assigning a higher score (Score 3) for responses with lexical deviations (e.g., lexical omissions or substitutions) and a lower score (Score 2) for responses with grammatical errors (e.g., structural omissions or incorrect prepositions).

In the following, we present a summary of the scoring rubric. Its use is exemplified for item #2 in Table 1. The complete scoring rubric can be found in the Supplementary Material to this paper.

Score 4 The response matches the stimulus sentence exactly or 1–2 typos are present.[2]

Score 3 Changes in grammar or lexical changes that preserve the original structure and result in grammatically correct and meaningful sentences, e.g. confusing definite and indefinite articles (where interchangeable), or (near-)synonymic substitutions of words.

Score 2 Changes in grammar that result in ungrammatical sentences or grammatical sentences which are not meaningful, e.g. violated agreement

[2]Typos include: transposed letters (all present), one letter replaced by a QWERTZ-adjacent key, one letter added/omitted next to an adjacent key, or a missing space between words.

Score	Example	Explanation
0	Bein praktikum	less than half of the words repeated correctly
1	Bei einem Praktikum * * *	half of the words repeated correctly but most of the meaning lost
2	Bei ein Praktikum lernt man viel.	case wrongly marked, ungrammatical sentence
3	Beim Praktikum lernt man viel.	change in grammar (contraction of preposition + article) but still grammatical
4	Bei einem Praktikum lenrt man viel.	one typo

Table 1: Example of the scoring rubric for the stimulus sentence *Bei einem Praktikum lernt man viel.* 'At an internship, one learns a lot'.

between subject and verb, structural omissions or wrong plural formation.

Score 1 More than half of the words are repeated but a considerable part of the original meaning or structure is lost or changed.

Score 0 Less than half of the words are repeated.

Some score descriptors vary with the length of the stimulus sentence: in "shorter sentences" (≤ 15 syllables) fewer deviations are allowed than in "longer sentences" (> 15 syllables). If a response contains multiple deviations at different score levels, then the lowest score determines the overall score. An accumulation rule is applied when two or more deviations of the same level are present in scores 2 or 3, leading to an overall score of 1 or 2, respectively. Punctuation and capitalization of the first word of the sentence are not taken into account.

The gold standard scores (henceforth also referred to as 'gold scores') for our study were assigned by three human raters in the following procedure: First, they familiarized themselves with the assessment rubric and participated in a calibration session using 200 responses (i.e. for all 20 items a sample of 10 participants each). Following this, a sample of the same size was randomly selected for independent evaluation by each rater. The inter-rater reliability (Fleiss' κ) for the resulting 200 ratings averaged around .986, indicating almost perfect agreement, with values ranging from .895 to 1.0 across the 20 items. The remaining responses were rated by one rater each, and ratings were discussed by all raters throughout the process to address difficult cases and ensure consistency.

3.3 Data Splitting

Each of the 195 participants responded to 20 different stimulus sentences (*items*). In total, our dataset comprises 3,900 pairs of stimulus (*target*) and imitation (*response*) sentences. We split the data into training, validation, and two different test sets as follows: First, we set aside all responses to two of the stimulus sentences (#4 and #18, i.e. one ≤ 15 syllables and one > 15 syllables, see Section 3.2). We call this test set 'Test unknown', comprising 390 stimulus-response pairs in total. The rest of the data was randomly split into a training, validation, and another test set. We call this second test set 'Test known', because it contains responses to those stimulus sentences that are also part of the training set. By using these two different test sets, we are able to not only assess how well a model performs on unseen response sentences to known stimulus sentences but also how well it generalizes to completely new stimulus sentences. The resulting data distribution across sets and gold scores is shown in Table 2.

Score	Train	Val.	Test known	Test unk.	Total
0	1,095	25	25	87	1,232 (32%)
1	701	25	25	92	843 (22%)
2	553	25	25	62	665 (17%)
3	260	25	25	78	388 (10%)
4	651	25	25	71	772 (20%)
Total	3,260	125	125	390	3,900 (100%)

Table 2: Number of stimulus-response pairs in the training, validation, and test sets, respectively, per gold score. 'Test unk.' contains stimulus sentences held out from the training set, 'Test known' a random subset of the remaining data (i.e. stimulus sentences known in the training set).

4 Method

We first present our deep learning model (**DL model**) and then introduce the pipeline for the rule-based model (**RB model**) for automatically scoring the WEIT.

4.1 Deep Learning

Since there is not enough data to train a deep learning model from scratch, we decided to use a pre-trained transformer model and fine-tune it on our

data for multi-class sequence classification.

For efficiency reasons, we chose the DistilBERT model (Sanh et al., 2019), a distilled version of BERT (Devlin et al., 2019), which the authors showed to be 40% smaller, 60% faster and able to retain 97% of the original language understanding capabilities (Sanh et al., 2019). We used the pre-trained model for German cased data (distilbert-base-german-cased) with a multi-label classification head, and fine-tuned the model on our WEIT training set. Hyper-parameters were optimized based on accuracy on the validation set, yielding the following setup and parameter values: a learning rate of 1e-5 and an epsilon value of 1.5e-3 for the Adam optimizer, and the default loss function for multi-class classification (SparseCategoricalCrossentropy). We trained the model for 50 epochs with an early stopping mechanism triggered after 5 consecutive epochs without improvement in the validation loss. The training and validation data were shuffled and batched in each iteration, with a batch size of 16.

Since the training dataset was heavily skewed towards scores 0, 1, and 4 (see Table 2), we trained a second model in the same way but in which class weights were introduced for scores 2 and 3. Score 2 received a 2x multiplier and score 3 received a 4.5x multiplier, both approximately equal to the proportion of the corresponding training pairs of these scores to the number of score 0 pairs (the most common score). We refer to this as the **weighted** DL model and the model without adjusted class weights as the **unweighted** DL model.

4.2 Rule-Based

The rule-based model processes pairs of target and response sentences through a multi-step pipeline to generate a score (Figure 1).

Preprocessing In the preprocessing step, the sentences are normalized and cleaned so that differences between target and response sentences that are not relevant for scoring can be ignored. This means in particular: capitalizing the first letters of both sentences, transforming common umlaut variants into the correct character (e.g., 'ae' into 'ä'), and removing punctuation and artifacts such as the ';timeout' token, which appears when a participant runs out of time during the repetition process. Furthermore, in some cases participants repeated the sentence multiple times. Since this is ignored by the human raters, we cut each response to only retain the first response sentence.

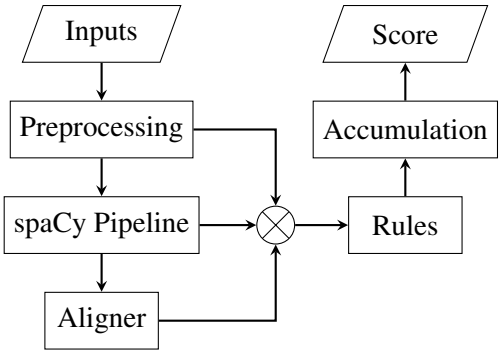

Figure 1: Flow diagram illustrating the rule-based model's data processing pipeline.

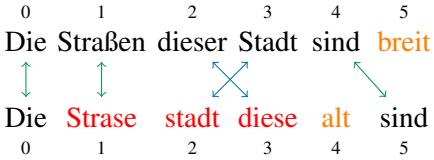

Figure 2: Token mapping by the aligner function for an example sentence. Tokens in red are **misspelled** and tokens in orange are **missing** or **additional**. Green arrows denote aligned tokens and blue arrows **transpositions**.

Linguistic Annotation In the next step, the preprocessed sentences are analyzed linguistically using a spaCy pipeline (Honnibal et al., 2020), which transforms each sentence into a list of tokens with part-of-speech (POS) tags, syntactic dependency labels, morphological features, and syllable counts.[3]

Alignment The tokens are then passed on to a custom-built aligner, which maps the words in the response to the words in the target sentence and also detects missing or added words (see Figure 2). This is done by calculating a matrix of Damerau-Levenshtein distances[4] between all words in the response sentence and the target sentence and aligning those words with the smallest distance. We do not only align identical words because this would prevent misspelled words from being matched with the correct word in the target sentence. However, if the edit distance between two words is large, it is more likely a different word rather than a misspelling. Therefore, for two words to be aligned,

[3]We use spaCy v3.8.3 with the de_core_news_sm model v3.8.0 with all its default components, and the package *sloev/spacy-syllables* v3.0.2 for counting the syllables, which is added directly after the tagger in the spaCy processing pipeline.

[4]using *lanl/pyxDamerauLevenshtein* v1.8.0, `https://github.com/lanl/pyxDamerauLevenshtein`)

their edit distance must be ≤ 3.[5] If a word in the target sentence has no match in the response sentence, it is '**missing**', if a word in the response sentence has no match in the target sentence, it is considered '**additional**'. Note that at this step, we do not detect word substitutions directly but they would be treated as a missing and an additional token, which are later aligned by a rule that checks for substitutions. For all matched pairs, if the edit distance between both words is greater than zero, the token in the response is considered '**misspelled**'. If two words are matched but have different positions, they are considered '**transposed**'. All other tokens are labeled as '**correct**'.

Rule Application Manually, a set of rules was crafted that implement the deviation categories from the scoring rubric based on all the outputs of the previous steps. For each category it is checked whether it applies to the response sentence. For this first version of the rule-based model, most deviation categories were implemented, except for some which were considered too fuzzy or which would have required further linguistic annotation not readily available e.g. about German plural formation.[6] The rules are defined in a way that they are mutually exclusive so that the order in which they are applied is not important. If a rule detects that a particular category applies to a response sentence, it outputs the name of the category, the score which it is associated with and how many instances of this deviation are found. Finally, an accumulation function collects the outputs of all rules and calculates the final score (see Section 3.2 for the accumulation rules). The following examples illustrate how some of the categories from the scoring rubric are approximated via rules.

To detect an *Omission Error*, the rule uses the missing-word count from the aligner. If exactly one word is missing, the rule assigns a score of 3. If two words are missing in a sentence with fewer than 16 syllables, the score is 2. In longer sentences with two or more omissions, the rubric asks to assess whether the sentence "preserves most of the original sentence structure and most of the meaning". We determine structural deviations by the degree of agreement between the spaCy dependency structures of stimulus and response sentence, with a loss of more than 30% of the original dependencies serving as the threshold. Meaning deviations are identified using cosine similarity between the vectorized representations of target and response sentences, obtained via a BERT Sentence Transformer (Reimers and Gurevych, 2019). If the similarity falls below 0.987, the meaning is considered altered.[7] If either a structural or meaning deviation occurs, the score is set to 1, otherwise 2.

The *Changes in Grammar* category captures deviations in grammatical structure between the stimulus and response sentences, which are specifically listed in the scoring rubric, namely differences in article usage, gender and case markings, agreement violations, and prepositional errors. The rule uses information from the aligner and spaCy to compare the POS and morphological features of aligned words. Article-related errors are identified when a determiner is missing, incorrectly added, or replaced with another. Gender and case errors are identified when mismatches occur in the morphological features of aligned words. Agreement violations are detected by comparing the number feature between a verb and its subject. Finally, prepositional errors include missing or incorrect prepositions. The scoring mechanism assigns a score of 2 for each error, counting the number of detected grammatical mistakes to determine the final score.

5 Evaluation

We evaluate the weighted DL model, the unweighted DL model and the RB model on the test set with known items and unknown items, respectively, as well as on the combination of the two test sets (henceforth called combined test set). Table 3 reports the accuracy, i.e. how often the exact gold score was predicted, and Quadratically Weighted Kappa (QWK), which penalizes greater deviations from the gold score more severely than smaller deviations.

For the DL models, we expected a drop in performance when comparing the scoring of responses to known versus unknown items, but not for the RB model. In fact, we see a considerable drop for the DL models: For example, QWK decreases from .93 to .62 for the unweighted DL model and from

[5] This value worked well in our trial runs but could be tuned, e.g. adjusted for token length, in future work.

[6] Deviation categories that were not implemented are: *wrong plural formation*, *missing structural elements or wrong word order*, and *sentence is grammatical but not meaningful* from score 2, and *changes in grammar that preserve the original structure and result in grammatically correct sentences* from score 3.

[7] The thresholds worked well in our trial runs but could be tuned more systematically in future work.

Model	Known It.		Unknown It.		Combined	
	acc	qwk	acc	qwk	acc	qwk
DL unw.	.71	.93	.46	.62	.52	.72
DL weigh.	**.78**	**.94**	.51	.83	.57	**.87**
RB	.73	.90	**.66**	**.87**	**.68**	**.87**

Table 3: Accuracy and QWK for the deep learning models (DL) without class weights ('unw.'), with class weights ('weigh.') and the rule-based model (RB), respectively, on the test sets with known items and unknown items, respectively, and the combined test set. Numbers in bold indicate the best model per set and metric.

DL	RB	Gold	Count	Perc.
•	•	•	209	41%
•	-	•	87	17%
-	•	•	140	27%
•	•	-	53	10%
-	-	-	26	5%
Total			515	100%

Table 4: Number of responses in the combined test set for which all three, only two or none of the scores given by the deep learning model (DL), the rule-based model (RB) and the gold standard are identical. '•' indicates that the same score was assigned.

.94 to .83 for the weighted DL model. For the RB model, there is only a slight drop from .90 to .87, which may also be due to chance considering the rather small test sets.

Overall, the weighted DL model is the best performing model on the known items, while the RB model is the best performing model on unknown items. On the combined test set, both models perform on par in terms of QWK (.87), but the RB model attains higher accuracy (.68 compared to .57). The weighted DL model consistently outperforms the unweighted DL model.

When looking at the confusion matrices of the three models based on the combined test set (Figure 3), we see that the greatest weakness of the unweighted DL model is that it hardly ever predicts score 3 and only rarely score 4. In fact, on unknown items it never predicts score 3 and only once score 4, hence it fails to generalize when a response is to be counted as (almost) correct. For the other two models, we see almost no large deviations from the gold standard, which was already reflected in the overall high QWK scores.[8]

5.1 Fine-Grained Model Comparison

In the following, we will restrict the discussion to the weighted DL model and the RB model and look more closely into their commonalities and differences.

From the confusion matrices (Figure 3) we see that the RB model has a distinct tendency to undervalue the responses: Out of 176 misclassified responses, 147 (84%) receive a score lower than the gold standard, i.e. the model tends to be stricter than the human raters. For the DL model, there is a similar trend, but proportionally it is not as extreme: Out of 219 misclassified responses, 157 (72%) are undervalued (note that in terms of absolute numbers, there are more undervalued items for the DL than for the RB model).

In Table 3, we saw that the RB model had an overall higher accuracy on the combined test set than the DL model. But does this mean that it correctly predicts most of the responses that the DL model also scores correctly – plus some additional ones – or do the two models actually succeed on different sets of responses?

Table 4 shows a breakdown of how often either both models or only one of them or none agrees with the gold standard and how often the two models agree with each other on the combined test set. In sum, only for 51% of the responses, the DL model and the RB model predict the same score. When they agree with each other, this does not necessarily mean that they are correct because for 10% of the responses, both models agree but they both deviate from the human gold standard (which can, in fact, also point to human ratings being inconsistent with the scoring rubric, see Section 5.2). On the other hand, for 85% of the responses at least one of the models is correct, i.e. agrees with the human gold standard. For 27%, only the RB model is correct and for 17% only the DL model. This indicates that both models have different strengths and weaknesses that we will examine more closely in the following.

Table 5 shows a breakdown of precision, recall, and F1-score per gold score for each of the two models. We see that for all scores but score 2, the RB model performs better or on par with the DL model. For score 4, the difference is most striking, with a very high recall of the RB model (.96) and

[8]There is one extreme outlier where the RB model predicts score 0, while the gold score is 4. This occurred because the response contained multiple repetitions of the stimulus sentence, and a bug prevented truncation of this particular case, contrary to what was prescribed by the preprocessing step described in Section 4.2.

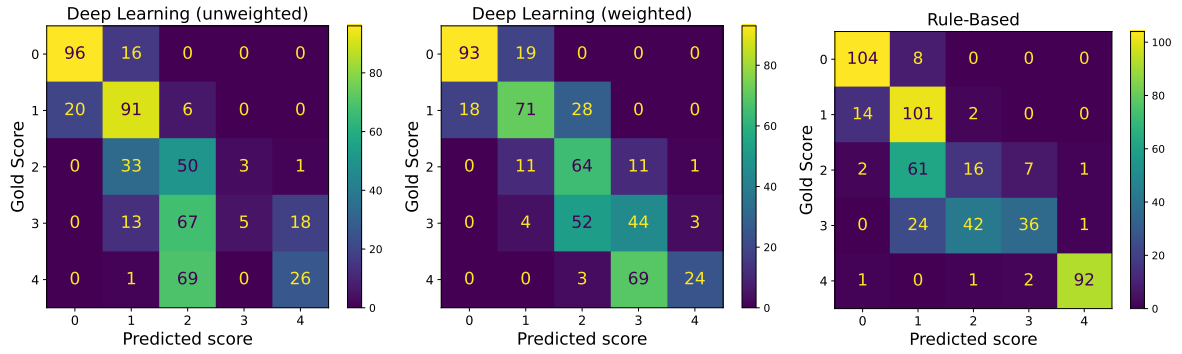

Figure 3: Confusion matrix of gold score vs. predicted score per model.

Score	Precision		Recall		F1	
	DL	RB	DL	RB	DL	RB
0	.84	**.86**	.83	**.93**	.83	**.89**
1	**.68**	.52	.61	**.86**	.64	**.65**
2	**.44**	.26	**.74**	.18	**.55**	.22
3	.35	**.80**	**.43**	.35	.39	**.49**
4	.86	**.98**	.25	**.96**	.39	**.97**
macro avg	.63	**.68**	.57	**.66**	.56	**.64**
micro avg	.64	**.69**	.57	**.68**	.57	**.66**

Table 5: Precision, recall, and F1-score per gold score, as well as macro average and micro (=weighted) average, for the deep learning model with class weights (DL) and the rule-based model (RB) based on the combined test set. The numbers in bold indicate the higher value in precision, recall, and F1-score, respectively, per score.

a very low recall of the DL model (.25). Only for score 2, the DL model clearly outperforms the RB model. This is partly in line with our expectation that the DL model performs better at scores where the scoring rubric categories are rather vague (e.g., responses with changes in grammar can receive either score 2 or score 3 depending on whether the sentence is still grammatical and meaningful). We will qualitatively discuss some of the misclassifications in the following section.

5.2 Discussion of Misclassifications

In the following, we will qualitatively discuss some of the misclassifications of the models to identify their potential limits and also to find leverage points for improvement.

Limitations of the deep learning model We saw that the deep learning model has a strikingly low recall for score 4. In fact, except for one response, the cases where the model failed to predict score 4 were caused by responses to the two previously unknown items. This indicates that the model failed to generalize to new sentences when a response is to be rated as fully correct. This is the case even when the responses are exact repetitions of the target sentence (38 out of 72 misclassifications). While these misclassifications could potentially be eliminated by passing a similarity score to the model, the remaining errors are harder to mitigate. This concerns, for example, accepted typos in a response, where the advantage of the RB model is that we can specify exactly what counts as a typo.

Limitations of the rule-based model For the DL model, we do not easily know why a response was misclassified but for the RB model we can analyze which categories were missed or falsely detected. We found some systematic causes for misclassifications:

Firstly, the model sometimes fails to differentiate between spelling errors, typos, and grammatical errors. One particular problem is the treatment of real-word spelling errors, i.e. (potential) spelling errors that result in another existing word form, e.g. *fährt/fahrt* (3SG/2PL of '(to) drive' or *es/er* ('it/he'). They make the sentence ungrammatical or not meaningful but are overvalued by the model because only a spelling error is detected. On the other hand, misspellings can result in nonsense words that are unknown to spaCy, which impacts the syntactic or morphological analysis of the sentence. For example, we found that when the spelling of *Musik* ('music') is changed to *Music*, spaCy assigns it neuter gender (instead of feminine), so that a model classifies the sentence as containing a grammatical error.

Furthermore, the model cannot determine well whether a substitution preserves the overall meaning and grammatical structure of the sentence, leading to an undervaluation of examples like *Die*

*Häuser sind nicht **sehr/so** schön* ('The houses are not **very/so** pretty') or *Kosten **der** Häuser / Kosten **von** Häusern* ('costs of the houses'). This is mainly important to differentiate between scores 2 and 3, which explains the low performance of the model for these scores.

Limitations of the human ratings In fact, not all deviations from the gold standard turned out to be true misclassifications. In some cases the models uncovered inconsistencies in human ratings, e.g. where human raters had overlooked deviations or not followed the rubric, but these cases were rare.

6 Conclusion and Future Work

We implemented an automated scoring procedure of a German WEIT using two approaches: a rule-based approach with manually crafted rules implementing the specific categories listed in the scoring rubric, and a deep learning approach that received pairs of stimulus sentences and test-takers' responses as training data. We found that the overall performance of both kinds of models is promising but not yet optimal, and that both approaches have different strengths and weaknesses. The rule-based model outperformed the deep learning model on previously unseen stimulus sentences and for the scores at the edges of the rating scale. The deep learning model, in contrast, was more successful in some cases of mid-range scores, for which explicit rules are harder to define.

The results indicate that a promising direction for future research could be to develop an ensemble or hybrid model: using rule-based scoring for categories with very high precision, and training a DL model only for those where clear rules are difficult to define. It also remains to be investigated whether Large Language Models (LLMs) with their broad language comprehension capabilities could contribute to the automated scoring or detection of specific error categories.

Limitations

One clear limitation of our study is that we only evaluated one deep learning model (DistilBERT). Different models, especially models operating on the character level, may lead to better results, e.g. by better capturing spelling errors and typos, and are worth investigating in future work. Furthermore, given that the class weighting had a great impact on the DL model, finding the optimal weighting could be investigated more systematically. In general, there could be a more systematic fine-tuning of hyperparameters but this would require access to larger computational resources, e.g. servers, that we wanted to avoid. Furthermore, we only used spaCy and no other tools for annotating the linguistic structure of the responses, which has a considerable impact on the overall performance of the rule-based model. Trying other or combining different linguistic processing tools could improve the results. Another limitation is that some of the deviation categories from the scoring rubric pertaining to scores 2 and 3 were not implemented yet in the rule-based model, which probably in part accounts for the weaker performance of the model for these scores compared to the other scores. Some of these rules could be implemented in future work by adding further specific resources (e.g. about German plural formation) while others, such as detecting sentences which are grammatical but not meaningful, could be tackled by using LLMs or finetuning models to specifically detect such cases.

Ethical Considerations

Our study investigated whether it is, in principle, feasible to automatically score a German WEIT. Our aim was to approximate human ratings as closely as possible, which means that there is a risk that potential biases in human ratings could be inherited by automated scoring systems. Furthermore, any biases present in the dataset may be reflected in the models. If the automated scoring of the test was used in a real-world application, it could have positive ethical impacts such as a better accessibility of language tests where they would otherwise not be available due to a lack of human raters. However, in a real-world scenario, a range of further ethical considerations would apply, e.g. regarding fairness, whose discussion is beyond the scope of this paper.

Acknowledgments

The WEIT for German was conceived, developed, and the data collected as part of a research project funded by the Deutsche Forschungsgemeinschaft (DFG), grant number 462766474. We would also like to thank the anonymous reviewers for their very helpful comments.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Anastasia Drackert. 2016. *Validating Language Proficiency Assessments in Second Language Acquisition Research. Applying an Argument-Based Approach*. Peter Lang.

C. Ray Graham, Deryle Lonsdale, Casey Kennington, Aaron Johnson, and Jeremiah McGhee. 2008. Elicited imitation as an oral proficiency measure with ASR scoring. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Daniel R. Isbell, Kathy MinHye Kim, and Xiaobin Chen. 2023. Exploring the potential of automated speech recognition for scoring the Korean Elicited Imitation Test. *Research Methods in Applied Linguistics*, 2(3):100076.

Maria Kostromitina and Luke Plonsky. 2022. Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, 44(3):886–911.

Deryle Lonsdale and Carl Christensen. 2011. Automating the scoring of elicited imitation tests. In *Proc. Machine Learning in Speech and Language Processing (MLSLP 2011)*, pages 16–20.

Michael McGuire and Jenifer Larson-Hall. 2025. Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: Steps toward automation. *Research Methods in Applied Linguistics*, 4(1):100197.

Benjamin J Millard. 2011. Oral Proficiency Assessment of French Using an Elicited Imitation Test and Automatic Speech Recognition. Master's thesis, Brigham Young University.

Majid Nikouee and Leila Ranta. 2023. Building an elicited imitation task as a measure of implicit grammatical knowledge. *Instructed Second Language Acquisition*, 7(1):41–67.

Lourdes Ortega, Noriko Iwashita, John M Norris, and Sara Rabie. 2002. An investigation of elicited imitation tasks in crosslinguistic SLA research. In *Second Language Research Forum, Toronto*, pages 3–6.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Hui Sun, Dagmar Divjak, and Petar Milin. 2025. Introducing fluency measures to the elicited imitation task. *Research Methods in Applied Linguistics*, 4(1):100176.

Anna Timukova, Oleksandra Yazdanfar, and Anastasia Drackert. submitted. So viele Lücken, so wenig Zeit: Die Rolle der Zeit im Konstrukt des deutschen C-Tests anhand der Analyse der Verarbeitungsprozesse (So many gaps, so little time: The role of time in the construct of the German C-Test based on the analysis of response processes). *Zeitschrift für Interkulturellen Fremdsprachenunterricht (ZIF) (Journal for Intercultural Foreign Lanugage Teaching)*.

Xun Yan, Yukiko Maeda, Jing Lv, and April Ginther. 2016. Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing*, 33(4):497–528.

A List of Items

#	Item	# Syllables
1	Die Straßen dieser Stadt sind breit. *The streets of this city are wide.*	8
2	Bei einem Praktikum lernt man viel. *At an internship, one learns a lot.*	9
3	Ich glaube nicht, dass er gut fahren kann. *I don't think that he can drive well.*	10
4	Die Häuser sind nicht sehr schön und viel zu teuer. *The houses are not very nice and far too expensive.*	12
5	Der Junge, dessen Katze gestern starb, ist traurig. *The boy whose cat died yesterday is sad.*	13
6	Das Restaurant sollte sehr gutes Essen haben. *The restaurant should have very good food.*	13
7	Du magst es sehr gerne, alte Musik anzuhören. *You like it a lot to listen to old music.*	14
8	Sie hat vor Kurzem ihre Wohnung fertig gestrichen. *She recently finished painting her apartment.*	14
9	Sie bestellt immer nur Fleisch und isst gar kein Gemüse. *She only ever orders meat and doesn't eat any vegetables.*	14
10	Meine Ehefrau hat einen sehr guten Sinn für Humor. *My wife has a very good sense of humor.*	15
11	Den meisten Spaß hatte ich als wir in der Oper waren. *I had the most fun when we were at the opera.*	15
12	Ich wünschte, dass ich mir die Kosten von Häusern leisten könnte. *I wish I could afford the cost of houses.*	16
13	Ich hoffe, dass es dieses Jahr früher wärmer wird als letztes. *I hope it gets warmer earlier this year than last.*	16
14	Bevor er nach draußen gehen kann, muss er sein Zimmer aufräumen. *Before he can go outside, he has to tidy his room.*	17
15	Ein Freund von mir passt immer auf die drei Kinder meines Nachbarn auf. *A friend of mine always looks after my neighbor's three children.*	17
16	Die Prüfung war nicht so schwer im Vergleich zu dem was Du mir erzählt hast. *The exam wasn't that difficult compared to what you told me.*	18
17	Die Anzahl von Leuten, die Zigaretten rauchen, steigt doch jedes Jahr mehr. *The number of people who smoke cigarettes is increasing every year.*	19
18	Je kleiner eine Universität ist, desto besser ist die Betreuung. *The smaller the university, the better the support.*	20
19	Wie in vielen europäischen Ländern gibt es auch in Deutschland einen Mindestlohn. *As in many European countries, there is also a minimum wage in Germany.*	22
20	Eine Fremdsprache hat sowohl einen persönlichen als auch einen beruflichen Nutzen. *A foreign language has both personal and professional benefits.*	24

Table 6: Full list of items used in the WEIT. English translations in italics are only added for clarity here and are not part of the test.