



# **RISTAL 8/2025**

## **Research in Subject-matter Teaching and Learning**

### **Special Edition: Reading, writing and interpreting texts in digital environments**

**Edited by  
Stefan D. Keller & Volker Frederking**

**Citation:**

Strahl, F., Kilian, J. & Möller, J. (2025). Judging Students' Texts in a Digital Research Tool: Do Text Quality, Students' Gender, and Migration Background Impact Teachers' Text Assessments? RISTAL, 8(2), 47-69.

DOI: <https://doi.org/10.2478/ristal-2025-0006>

ISSN 1863-0502



© 2025 Frederike Strahl, Jörg Kilian & Jens Möller. This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

# Judging Students' Texts in a Digital Research Tool: Do Text Quality, Students' Gender, and Migration Background Impact Teachers' Text Assessments?

*Frederike Strahl\*, Jörg Kilian\*\* & Jens Möller\*\*\**

## Abstract

Teachers' assessments of students' writing are key for adapting instruction to individual needs but are often influenced by biases. The present studies examined whether student characteristics – gender and migration background – affect text assessments in the digital tool Student Inventory. In Study 1 (n = 117), pre-service teachers assessed six texts differing in quality and students' gender. In Study 2 (n = 127), students' migration background was manipulated instead. Texts were assessed holistically and analytically, considering content, style, and linguistic accuracy. Results from both studies showed that participants reliably differentiated between low- and high-quality texts on all assessment scales. The studies revealed small significant gender biases only for medium-quality texts. The findings suggest that the Student Inventory facilitates an objective assessment of students' writing and supports fair assessment practices, regardless of students' gender or migration background.

**Keywords:** assessing writing, teachers' judgment, diagnostic competence, gender, migration background, assessment biases

---

## \* **Frederike Strahl**

is a doctoral researcher interested in teachers' diagnostic competence and judgment accuracy in education. Institute of Psychology of Learning and Instruction, Kiel University, Olshausenstraße 75, 24118 Kiel, Germany  
E-Mail: fstrahl@ipl.uni-kiel.de

## \*\* **Jörg Kilian**

is a full professor interested in German didactics. German Philology/Didactics of German Language at the Institute of German, Leibnizstraße 8, 24118 Kiel, Germany  
E-Mail: kilian@germsem.uni-kiel.de

## \*\*\* **Jens Möller**

is a full professor interested in learning and instruction. Institute of Psychology of Learning and Instruction, Kiel University, Olshausenstraße 75, 24118 Kiel, Germany, Olshausenstraße 75, D-24118 Kiel  
E-Mail: jmoeller@ipl.uni-kiel.de

## Author Contributions

**Frederike Strahl:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing – Original Draft, Visualization, Project administration. **Jörg Kilian:** Writing – Review & Editing, Project administration, Funding acquisition. **Jens Möller:** Conceptualization, Writing – Review & Editing, Supervision, Project administration, Funding acquisition.

Declaration of interest: none

## 1 Introduction

The accurate assessment of students' writing competencies is a prerequisite for adaptive instructional design in language subjects (Graham et al., 2015; Herppich et al., 2018). Teachers' assessment of students' written performances is central to this (Eckes, 2008; Feenstra, 2014). However, studies have shown low agreement among teachers assessing the same text (Cooksey et al., 2007; Jansen, Vögelin, Machts, Keller, Köller, & Möller, 2021; Möller et al., 2022; Skar & Jølle, 2017), suggesting that performance-irrelevant factors – such as gender or migration background – may influence teachers' assessments (Kaiser et al., 2017; Karing, 2009; Stang & Urhahne, 2016). These characteristics are often associated with stereotypes in school contexts (Copur-Gencturk et al., 2023; Retelsdorf et al., 2015). Girls tend to receive higher grades in language subjects (Hartig & Jude, 2008; Petersen, 2018; Reilly et al., 2019), and students with a migration background tend to receive lower results for their performances in general (Gebhardt et al., 2013). Experimental research controlling for students' performance levels found that gender and migration background influenced teachers' assessments (Bonefeld et al., 2022; Ready & Wright, 2011), although findings are inconsistent: some studies found no such effects (Copur-Gencturk et al., 2022, 2023; Doornkamp et al., 2022; Karing et al., 2024).

Text assessments may be particularly prone to bias due to their complexity and multidimensionality, even when clear criteria are provided (Cooksey et al., 2007; Jansen, Vögelin, Machts, Keller, Köller, & Möller, 2021; Möller et al., 2022; Skar & Jølle, 2017). Several studies investigated biases in teachers' performance assessment in digital settings (Copur-Gencturk et al., 2022; Jansen, Vögelin, Machts, Keller, Köller, & Möller, 2021; Möller et al., 2022; Strahl et al., 2025), but evidence of students' gender and migration background on first-language writing is lacking. Since first-language processing is often more automatic than second-language processing (Clahsen & Felser, 2006; Ullman, 2020), teachers may rely more heavily on readily available information, such as stereotypes, thereby increasing the likelihood of bias. To ensure an advantage of digital assessment tools in school praxis, it is necessary to investigate whether gender and migration background biases persist in online assessment environments. The present studies address these gaps by investigating whether German pre-service teachers show bias related to students' gender and migration background when assessing first-language texts in an online environment. Moreover, we explore the potential of the digital assessment tool Student Inventory not only for controlled experimentation, but also as a prospective training tool to raise awareness for judgment biases and foster diagnostic competencies in teacher education.

In sum, while prior studies have shown that teachers' assessments can be influenced by irrelevant student characteristics such as gender or migration background, robust evidence for these effects in digital environments and with first-language texts remains scarce. At the same time, digital tools offer promising opportunities not only for standardized assessment but also for improving teacher training by making judgment processes more transparent and bias-sensitive. The present studies aim to address both of these issues by using the Student Inventory to test for potential bias and to explore its utility as a tool for supporting accurate and fair assessment practices in teacher education.

### *1.1 Bias in Text Assessment: The Role of Student Characteristics*

Research on teachers' judgment accuracy highlights that student characteristics such as gender and migration background can influence teachers' performance assessments (see the heuristic model of judgment accuracy; Südkamp et al., 2012). According to the continuum model of impression formation (Fiske & Neuberg, 1990), performance-irrelevant cues can activate stereotypes, particularly in complex tasks such as text assessment. Empirical findings support that text assessment is vulnerable to such irrelevant factors, including text length (Barkaoui, 2010; Fleckenstein et al., 2020), teachers' experience (Möller et al., 2022), the performance level of reference texts (Strahl et al., 2025), and halo effects (Jansen, Vögelin, Machts, Keller, & Möller, 2021; Vögelin et al., 2018, 2019). Loibl et al.'s DiaCom model (2020) also emphasizes that student characteristics may activate internal beliefs and stereotypes, which in turn shape teachers' assessment decisions.

For instance, gender stereotypes can lead teachers to assume that girls are more competent in language-related tasks, often resulting in more favorable assessments in language subjects (e.g., Gentrup et al., 2024; Ready & Wright, 2011), while boys are more positively rated in mathematics (e.g., Bonefeld et al., 2022). However, some studies do not find these effects (e.g., Copur-Gencturk et al., 2022), suggesting that gender bias is not universal.

Similarly, teachers may hold lower expectations for students with a migration background, possibly due to assumptions about their language proficiency or the level of educational support available at home. These expectations can result in lower performance ratings – even when actual performance is comparable (e.g., Holder & Kessels, 2017). However, other studies report no significant effects (e.g., van Ewijk, 2011).

Despite these mixed findings, only a few studies have examined these biases in the specific context of text assessment, which is both cognitively demanding and open to interpretation (e.g., Jansen et al., 2019). This highlights the need for further research on how and when such biases emerge in teachers' text assessments.

### *1.2 Digital Contexts in Teachers' Text Assessment*

Digital tools offer new opportunities for both standardized assessment and bias-sensitive teacher education. These tools can simulate assessment situations, providing a more standardized representation of information (Machts et al., 2023). This can enhance teachers' ability to focus on performance-relevant cues during their assessment (Brunswik, 1955) and help reduce their cognitive load.

Only a few studies have explored whether teachers' assessment biases persist in digital assessment settings (Copur-Gencturk et al., 2022; Jansen et al., 2019; Jansen, Vögelin, Machts, Keller, & Möller, 2021; Möller et al., 2022; Strahl et al., 2025). Copur-Gencturk et al. (2022) and Jansen et al. (2019)

text assessment. Therefore, we conducted two experimental studies using the online assessment tool Student Inventory (Kaiser et al., 2015). This tool enables controlled variation of influencing factors and has been used in various subjects, for example, in German (Möller et al., 2022; Strahl et al., 2025), English (Jansen et al., 2019; Jansen, Vögelin, Machts, Keller, & Möller, 2021), and Biology (Fischer et al., 2021).

In our studies, we examine whether the Student Inventory enables valid and fair assessments of students' texts independent of students' gender or migration background and whether it can serve as an effective training tool for pre-service teachers – enhancing their ability to make unbiased, performance-based assessments in authentic educational settings.

### **1.3 Present Studies**

The present experimental studies examine whether pre-service teachers' assessments of first-language texts are biased by students' gender or migration background and how well they distinguish texts of varying quality in an online environment with concrete assessment criteria. In both studies, participants assessed six German-language texts differing in quality (low, medium, high) and author identity, indicated by gendered or culturally coded names. Study 1 focused on gender (female vs. male names), Study 2 on migration background (coded vs. non-coded names). We expect pre-service teachers to reliably distinguish text quality levels and to assess texts more favorably when attributed to female students or those without a migration background.

## **2 Method**

### **2.1 Participants**

#### **2.1.1 Study 1**

The required sample size for repeated-measures ANOVAs (power = .95, medium effect size,  $\alpha < .05$ ) was calculated using G\*Power (Faul et al., 2007), resulting in a minimum of 34 participants. The final sample of 130 participants was reduced by 13 participants (12 were no pre-service teachers, and one had identical responses across all analytic assessment scales), resulting in a final sample of 117 pre-service teachers of German studies from a university in northern Germany. Participants were, on average, 24.56 years old ( $SD = 2.50$ ; range = 21–33), and 77.78 % were female. A total of 23.93 % were Bachelor's and 76.07 % Master's students in their third semester of studies ( $M = 3.25$ ,  $SD = 2.63$ ; the first semester of the Master's degree is considered the first). There is an optional offer at regular intervals but no compulsory course on text assessment for German pre-service teachers at this university. All pre-service teachers aspire to become secondary school teachers.

In terms of additional subjects, 88 participants studied social sciences, 33 language subjects, and five scientific subjects. German was the native language for 93.16 % of the participants, and none had prior experience with the rating procedures and scales.

### 2.1.2 Study 2

Using the same G\*Power parameters as in Study 1, the required sample size was again 34. The initial sample of 145 participants was reduced to 127 (17 were no pre-service teachers, and one had identical responses across all analytic assessment scales), resulting in a sample of 127 pre-service German teachers from the same university as in Study 1. The average age was 24.37 years ( $SD = 2.59$ ; range = 20–36), with 81.89 % of the participants being female. Participants included 25.20 % Bachelor's, 67.72 % Master's, and 7.09 % state examination students in their fourth semester of study ( $M = 4.00$ ,  $SD = 2.55$ ), with the first semester of a Master's degree considered the first semester. All pre-service teachers aspire to become secondary school teachers, and text assessment courses are just optional offers and not compulsory at this university.

Regarding additional subjects, 81 participants studied social sciences, 31 languages, and 18 sciences. German was the native language for 90.55 % of the participants.

### 2.2 Independent Variables: Gender, Migration Background and Text Quality

The studies examined the influence of the text writer's gender, migration background, and overall text quality on text assessment. Gender and migration background were indicated by the writer's name, shown above each student text. Across the six texts each participant assessed, names varied systematically. As the texts were authored by students in the ninth and tenth grades, common German names from 2007 were used, selected based on similar perceived attractiveness and intelligence (Rudolph et al., 2007). In Study 2, to reflect the most frequent migration background in Germany, three common Turkish female names from 2007 were used to ensure gender consistency.

Text quality was determined via expert ratings from the Institute for Educational Quality Improvement in Berlin (IQB; Canz et al., 2020). Two trained student raters from German or related fields assessed each text using a holistic scale (1 = worst rating to 5 = best rating) and the three analytic scales content, style, and linguistic accuracy (1 = worst rating to 4 = best rating). The scales are described in detail below. Raters received two to three days of training, including trial assessments (see Canz et al., 2020). Interrater reliability was moderate to substantial (ICC: .64 holistic, .66 content, .55 style, .67 linguistic accuracy; Landis & Koch, 1977). Expert ratings have been shown to significantly correlate with teacher judgments (Jansen et al., 2024).

### 2.3 Dependent Variable: Text Assessment

In both studies, the participants assessed each text on four assessment scales. The first was a holistic scale, ranging from 1 (worst rating) to 5 (best rating). The remaining three scales were analytical, focusing on content, style, and linguistic accuracy. These scales ranged from 1 (worst rating) to 4 (best rating). The rating scales developed by Schipolowski and Böhme (2016) were employed. Schipolowski and Böhme (2016) adapted the National Assessment of Educational Progress (NAEP) writing scales for their study (National Assessment Governing Board, 2011a, 2011b; National Center for Education Statistics, 2012). The translated scales are available in the supplementary material.

## 2.4 Material

The texts and expert ratings were provided by the IQB and stem from a norming study with 293 texts written by ninth- and tenth-grade students as part of a large-scale assessment in Germany (Canz et al., 2020). Students wrote short newspaper reports based on key points and were instructed to start with the most important information and include a headline.

For the studies, six texts were selected based on expert holistic ratings and text length. The studies were conducted using LimeSurvey (Version 3.28.77) via the Student Inventory. On each survey screen, the text appeared on the left, and the assessment form on the right. Additional information about each text was available on the same screen (see Figure S1 in the supplementary material).

## 2.5 Procedure

Participants completed the study on a tablet or computer. After providing informed consent, participants were informed about the voluntary nature of the study, pseudonymization of data, and their right to withdraw data within two weeks. Participants then received instructions about the procedure, the writing task, and the assessment process. The assessment scales – including detailed criteria for the analytical dimensions (content, style, and linguistic accuracy) – were explained in advance (see supplementary material). Subsequently, participants rated all six texts holistically. Then, they assessed the same texts on the three analytical scales. The order of the criteria presentation was fixed. Texts could be selected via a dropdown menu and switched at any time. Finally, participants completed a short demographic questionnaire.

## 2.6 Data analysis

Both studies used a 3 (text quality: high, medium, low)  $\times$  2 (gender: Female vs. male/migration background: yes vs. no) factorial design with repeated measures on both variables. For each assessment scale, a mixed ANOVA with post-hoc tests was conducted in RStudio (version 4.3.3). The alpha level was Bonferroni-corrected. Only complete datasets (i.e., with ratings on all scales) were analyzed. No outliers were detected; participants engaged fully with the scale ranges.

# 3 Results

## 3.1 Study 1

### 3.1.1 Descriptive Results

Table 1 presents the means and standard deviations of the text assessments across the four scales, categorized by text quality and student gender. As expected, high-quality texts received more favorable ratings than low-quality texts. Gender-related differences were minimal for high- and low-quality texts but more pronounced for medium-quality texts.

**Tab.1:** Means and standard deviations for all assessment scales by text quality and student gender

|                     |        | Text quality  |               |               |
|---------------------|--------|---------------|---------------|---------------|
|                     |        | Low           | Medium        | High          |
| Scale               | Gender | <i>M (SD)</i> | <i>M (SD)</i> | <i>M (SD)</i> |
| Holistic            | Male   | 2.88 (1.20)   | 2.80 (1.06)   | 3.38 (0.91)   |
|                     | Female | 2.80 (1.26)   | 3.24 (1.10)   | 3.44 (0.93)   |
| Content             | Male   | 2.55 (0.87)   | 2.68 (0.87)   | 2.76 (0.89)   |
|                     | Female | 2.55 (0.85)   | 2.96 (0.80)   | 2.87 (0.80)   |
| Style               | Male   | 2.44 (0.98)   | 2.50 (0.84)   | 2.86 (0.80)   |
|                     | Female | 2.43 (0.95)   | 2.80 (0.82)   | 2.86 (0.75)   |
| Linguistic Accuracy | Male   | 2.64 (1.05)   | 2.44 (1.09)   | 3.09 (0.68)   |
|                     | Female | 2.55 (1.07)   | 2.86 (1.04)   | 3.11 (0.65)   |

*Note.* The holistic scale ranged from 1 to 5, and the analytic scales (content, style, linguistic accuracy) ranged from 1 to 4.

### 3.1.2 Hypothesis Testing

The four ANOVAs revealed no main effect of text author gender on any assessment scale (holistic:  $F(1, 116) = 2.49, p = .12, \eta^2 = .004$ ; content:  $F(1, 116) = 3.68, p = .06, \eta^2 = .01$ ; style:  $F(1, 116) = 2.03, p = .16, \eta^2 = .003$ ; linguistic accuracy:  $F(1, 116) = 1.84, p = .18, \eta^2 = .004$ ), and no significant interactions between text quality and gender (holistic:  $F(1.84, 213.22) = 2.72, p = .07, \eta^2 = .01$ ; content:  $F(2, 232) = 1.53, p = .22, \eta^2 = .004$ ; style:  $F(1.57, 181.58) = 1.99, p = .15, \eta^2 = .01$ ; linguistic accuracy:  $F(1.69, 195.54) = 3.00, p = .06, \eta^2 = .01$ ). However, the  $p$ -value for holistic and linguistic assessments was below .10, and nearly all differences favored texts with female names. Exploratory paired  $t$ -tests for medium-quality texts revealed significant gender effects with a small effect size on all scales after Bonferroni-Holm correction, with female-named texts assessed more positively (for detailed results, see Table S1 in the supplementary material). No gender-related differences were found for low- or high-quality texts.

The main effects of text quality were significant on all assessment scales (holistic:  $F(2, 232) = 34.94, p < .001, \eta^2 = .05$ ; content:  $F(2, 232) = 15.35, p < .001, \eta^2 = .02$ ; style:  $F(2, 232) = 20.10, p < .001, \eta^2 = .04$ ; linguistic accuracy:  $F(2, 232) = 52.52, p < .001, \eta^2 = .06$ ). Post hoc comparisons showed higher ratings for high-quality versus low-quality texts across all scales ( $p < .001$ ). For style, medium-quality texts were rated higher than low-quality ( $p < .05$ ) and lower than high-quality texts ( $p < .01$ ). On the holistic and linguistic accuracy scale, medium- and low-quality texts were not distinguished, and no difference emerged in the content assessment between medium- and high-quality texts (see Table S2 for descriptive statistics and Table S3 for detailed post hoc results in the supplementary material). These results

indicate that pre-service teachers reliably differentiated between low- and high-quality texts but struggled to distinguish medium-quality texts.

## 3.2 Study 2

### 3.2.1 Descriptive Results

Table 2 presents means and standard deviations of text assessments across the four scales, categorized by text quality and student migration background. High-quality texts were rated more positively than low-quality texts – differences between names indicating a migration background and those that do not were minimal.

**Tab.2:** Means and standard deviations for all assessment scales by text quality and students' migration background

| Scale               | Migration Background | Text quality  |               |               |
|---------------------|----------------------|---------------|---------------|---------------|
|                     |                      | Low           | Medium        | High          |
|                     |                      | <i>M (SD)</i> | <i>M (SD)</i> | <i>M (SD)</i> |
| Holistic            | Without              | 2.72 (1.17)   | 2.96 (1.14)   | 3.41 (0.88)   |
|                     | With                 | 2.74 (1.14)   | 2.95 (1.12)   | 3.28 (0.86)   |
| Content             | Without              | 2.40 (0.81)   | 2.76 (0.90)   | 2.84 (0.75)   |
|                     | With                 | 2.47 (0.79)   | 2.68 (0.85)   | 2.74 (0.77)   |
| Style               | Without              | 2.37 (0.91)   | 2.70 (0.89)   | 2.86 (0.77)   |
|                     | With                 | 2.29 (0.89)   | 2.64 (0.70)   | 2.82 (0.75)   |
| Linguistic Accuracy | Without              | 2.61 (1.16)   | 2.62 (1.08)   | 3.09 (0.69)   |
|                     | With                 | 2.44 (1.07)   | 2.69 (1.01)   | 3.09 (0.71)   |

*Note.* The holistic scale ranged from 1 to 5, and the analytic scales (content, style, linguistic accuracy) ranged from 1 to 4.

### 3.2.2 Hypothesis Testing

There was no main effect of students' migration background on any assessment scale (holistic:  $F(1, 126) = 0.20, p = .65, \eta^2 = 0$ ; content:  $F(1, 126) = 0.40, p = .53, \eta^2 = .0005$ ; style:  $F(1, 126) = 0.89, p = .34, \eta^2 = .001$ ; linguistic accuracy:  $F(1, 126) = 0.15, p = .70, \eta^2 = .0003$ ). No interaction effects between text quality and migration background were found. Thus, there was no evidence of bias related to migration background in pre-service teachers' assessments. Text quality showed significant main effects across all assessment scales (holistic:  $F(1.87, 236.15) = 37.83, p < .001, \eta^2 = .06$ ; content:  $F(2, 252) = 20.88, p < .001, \eta^2 = .03$ ; style:  $F(2, 252) = 37.65, p < .001, \eta^2 = .06$ ; linguistic accuracy:  $F(2, 252) = 61.90, p < .001, \eta^2 = .06$ ). Post hoc tests confirmed significantly higher ratings for high-quality compared to low-quality texts across all scales ( $p < .001$ ). For style, medium-quality texts were rated higher than low-quality ( $p < .001$ ) and lower than high-quality texts ( $p = .04$ ). However, on holistic and linguistic scales, medium- and low-quality texts were not differentiated, and on the content scale, medium- and high-

quality texts received similar assessments (see Table S4 for descriptive statistics and Table S5 for detailed post hoc results in the supplementary material).

In sum, pre-service teachers clearly distinguished between low- and high-quality texts but showed limited sensitivity to medium-quality texts.

## 4 Discussion

The objective of the present studies was to investigate the extent to which pre-service teachers are capable of assessing the quality of first-language student texts using concrete assessment criteria and whether students' gender or migration background influences these assessments in the online tool Student Inventory. The interdisciplinary studies thereby promote both writing and educational psychology research on judgment accuracy.

Assessing students' texts is a time-consuming yet essential task in teachers' professional life (Crusan, 2010; Jansen Vögelin, Machts, Keller, & Möller, 2021; Vögelin et al., 2019; Weigle, 2007). Providing both summative and formative feedback is crucial for helping learners and allows teachers to adapt their instruction to students' needs (Urhahne & Wijnia, 2021). However, texts are complex and multidimensional constructs, making accurate assessment challenging (Graham, 2019). Therefore, it is important to explore the factors that influence teachers' assessment of texts.

In line with previous research (Copur-Gencturk et al., 2022; Doornkamp et al., 2022; Jansen et al., 2019), the present studies found no biases due to students' gender and migration background in pre-service teachers' first language text assessments using the online tool Student Inventory. However, exploratory post-hoc analyses of Study 1 revealed that medium-quality texts with a female name were assessed more positively than those with a male name, indicating that gender stereotypes influence teachers' text assessments, at least for medium-quality performances. However, the effect size found can only be classified as small, and for students' migration background, no such biases were observed. Subsequent studies should systematically examine the gender bias effect in the assessment of medium-quality texts using inferential methods. Furthermore, the results of both studies revealed that teachers can successfully distinguish between low-quality and high-quality texts but have difficulties differentiating medium-quality texts.

These findings imply that, in digital settings, pre-service teachers are generally not biased by students' names indicating a specific gender or migration background. Digital contexts appear to standardize assessment information and reduce the cognitive load of assessment tasks. This allows teachers to focus on cues that are relevant to the assessment.

However, the observed gender effect in medium-quality texts indicates that in ambiguous situations, assessment-irrelevant but easily accessible information, such as students' gender, is still used for simplification, even within the standardized context of digital tools. Pre-service teachers appear to incorporate stereotypes about gender differences into their assessments when they are insecure in the assessment situation, as they are with the medium-quality texts, which may simplify assessment tasks, as the continuum model of impression formation (Fiske & Neuberg, 1990) suggests. This finding raises exciting questions for future research: How accurate are teachers' judgments of medium-quality texts? Do they experience greater

uncertainty in these cases? Investigating such questions can help identify circumstances under which bias occurs and improve teacher training accordingly.

For teacher education, the results imply that teachers should be better prepared for the assessment of medium-quality students' performances. Many classroom texts fall into the medium-quality range. It is precisely in these ambiguous cases that biases may emerge. Teacher training should include diagnostic tasks with uncertain texts and promote the use of structured rubrics, guided peer discussions, and critical reflection on potential biases. Pre-service teachers should be made aware of the simplification process of stereotypes in their performance assessment. In addition, teacher training courses in German at universities must develop the professional competence of pre-service German teachers in didactically sound text assessment. To this end, a "didactic language criticism" module should be included as a compulsory part of the course curriculum. In this module, pre-service teachers can learn how they can promote the language education of school students in their profession as "language norm authorities" in the form of didactically adaptive text assessment.

The findings of Jansen et al. (2019), Copur-Gencturk et al. (2022), and the present studies suggest that biases associated with gender and migration background are less likely in a virtual assessment context. These results highlight the potential of online tools such as the Student Inventory to train teachers in text assessment. Beyond functioning as training environments, their digital nature allows for the standardized and anonymized presentation of texts, reducing contextual noise and enabling controlled variations of student characteristics. Such tools can sensitize pre-service teachers to their assumptions and support the development of reflective, criteria-based assessment skills.

#### **4.1 Limitations**

Due to the experimental study design, internal validity is ensured, but the results are not directly transferable to reality. In real classrooms, teachers assess a large number of student texts across a broader quality range and have access to additional contextual information, such as students' appearance, behavior, or prior achievement. Such information may activate biases more strongly than a name above a student text. Furthermore, in practice, teachers often develop their assessment criteria rather than being provided with them, which could further increase subjectivity. Since we found small gender biases in the assessment of medium-quality texts under controlled conditions, these effects may be even more pronounced in everyday classroom settings. The results refer only to the dimension of text assessment; transfer to other subjects would have to be investigated in order to rule out the possibility that gender bias is a language-related phenomenon.

However, the Student Inventory can be adapted to reflect more realistic conditions, especially for text assessment, offering access to a larger number of texts with varying quality levels, more student information (e.g., pictures, prior achievements, motivation), and adaptable assessment criteria. This flexibility enables its use for research and practice-oriented teacher training alike.

## 4.2 Implications for Practice

Despite the limitations, the present studies showed that the Student Inventory could be a valuable tool for training teachers' first- and second-language text assessment. Given that many pre-service teachers feel unprepared for text assessment after their studies (Rauin & Meier, 2007), the implementation of the Student Inventory in teacher education appears promising.

The tool enables pre-service teachers to assess a wide range of authentic texts and compare their assessments with expert ratings – without requiring additional effort from instructors. By systematically varying text and student characteristics (e.g., names, images, or performance indicators), it promotes awareness of bias and supports criteria-based assessment.

Moreover, the Student Inventory has already been used in various subjects, including Biology (Fischer et al., 2021), English as a second language (Jansen et al., 2019; Jansen, Vögelin, Machts, Keller, & Möller, 2021), and German as a first language (Möller et al., 2022; Strahl et al., 2025). Its flexibility allows it to be embedded in various parts of the teacher education curriculum and can expand the assessment practices of pre-service teachers during their educational training. Tools such as the Student Inventory should be used more widely in mandatory university courses to provide hands-on practice in text or other performance assessments and foster critical reflection on fairness and bias in the classroom.

## 4.3 Conclusion

The present studies showed that pre-service teachers could generally distinguish between low-quality and high-quality student texts and showed no systematic gender or migration background biases using the Student Inventory, consistent with Jansen et al. (2019). However, by differentiating three levels of text quality, this study revealed difficulties in assessing medium-quality texts and small gender-related biases in these cases. These findings highlight the need for targeted training to address ambiguous performance levels and support the use of digital tools like the Student Inventory to enhance assessment competence and bias awareness in teacher education.

## 5 Acknowledgements

This work was supported by the German Research Foundation (DFG) [grant numbers MO 648/30-1, KI 828/6-1].

## References

- Barkaoui, K. (2010). Explaining ESL essay holistic scores: A multilevel modeling approach. *Language Testing*, 27(4), 515–535.  
<https://doi.org/10.1177/0265532210368717>
- Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, 62(3), 193–217.  
<https://doi.org/10.1037/h0047470>
- Bonefeld, M., Kleen, H., & Glock, S. (2022). The effect of the interplay of gender and ethnicity on teachers judgements: Does the school subject matter? *The Journal of Experimental Education*, 90(4), 818–838.  
<https://doi.org/10.1080/00220973.2021.1878991>
- Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Assessing Writing*, 45, 100470.  
<https://doi.org/10.1016/j.asw.2020.100470>
- Clahsen, H., & Felser, C. (2006). How native-like is non-native language processing? *Trends in Cognitive Sciences*, 10(12), 564–570.  
<https://doi.org/10.1016/j.tics.2006.10.002>
- Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation*, 13(5), 401–434.  
<https://doi.org/10.1080/13803610701728311>
- Copur-Gencturk, Y., Thacker, I., & Cimpian, J. R. (2022). Teacher bias in the virtual classroom. *Computers & Education*, 191, 104627.  
<https://doi.org/10.1016/j.compedu.2022.104627>
- Copur-Gencturk, Y., Thacker, I., & Cimpian, J. R. (2023). Teachers' race and gender biases and the moderating effects of their beliefs and dispositions. *International Journal of STEM Education*, 10(1), 1–25.

- Gentrup, S., Olczyk, M., & Lorenz, G. (2024). Teacher stereotypes and teacher expectations at the intersection of student gender and socioeconomic status. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 56(1-2), 87–102.  
<https://doi.org/10.1026/0049-8637/a000291>
- Graham, S. (2019). Changing How Writing Is Taught. *Review of Research in Education*, 43(1), 277–303.  
<https://doi.org/10.3102/0091732X18821125>
- Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta analysis. *The Elementary School Journal*, 115(4), 523–547.  
<https://doi.org/10.1086/681947>
- Hartig, J., & Jude, N. (2008). Sprachkompetenzen von Mädchen und Jungen [Language competencies of girls and boys]. In E. Klieme (Ed.), *Beltz Pädagogik, Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* (pp. 202–207). Weinheim: Beltz.  
<https://doi.org/10.25656/01:3154>
- Herppich, S., Praetorius, A.-K., Förster, N., Karst, K., Leutner, D., Behrmann, L. et al. (2018). Teachers' assessment competence. Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*, 76, 181–193.  
<https://doi.org/10.1016/j.tate.2017.12.001>
- Holder, K., & Kessels, U. (2017). Gender and ethnic stereotypes in student teachers' judgments: A new look from a shifting standards perspective. *Social Psychology of Education*, 20(3), 471–490.  
<https://doi.org/10.1007/s11218-017-9384-z>
- Jansen, T., Meyer, J., Schipolowski, S., & Möller, J. (2024). Feedback on teachers' text assessment: Does it foster assessment accuracy and motivation? *Zeitschrift für Pädagogische Psychologie*, 38(1-2), 35–47.  
<https://doi.org/10.1024/1010-0652/a000365>
- Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., & Möller, J. (

- Loibl, K., Leuders, T., & Dörfler, T. (2020). A framework for explaining teachers' diagnostic judgements by cognitive modeling (DiaCoM). *Teaching and Teacher Education*, 91, 103059. <https://doi.org/10.1016/j.tate.2020.103059>
- Machts, N., Chernikova, O., Jansen, T., Weidenbusch, M., Fischer, F., & Möller, J. (2024). Categorization of simulated diagnostic situations and the salience of diagnostic information: Conceptual framework. *Zeitschrift für Pädagogische Psychologie*, 38(1–2), 3–13. <https://doi.org/10.1024/1010-0652/a000364>
- Möller, J., Jansen, T., Fleckenstein, J., Machts, N., Meyer, J., & Reble, R. (2022). Judgment accuracy of German student texts: Do teacher experience and content knowledge matter? *Teaching and Teacher Education*, 119, 103879. <https://doi.org/10.1016/j.tate.2022.103879>
- National Assessment Governing Board (2011a). *Developing achievement levels on the national assessment of educational progress for writing grades 8 and 12 in 2011 and grade 4 in 2013*. NAEP Writing ALS Design Document.
- National Assessment Governing Board (2011b). *Writing framework for the 2011 national assessment of educational progress*. U.S. Department of Education, Washington, D.C.
- National Center for Education Statistics (2012). *The nation's report card: Writing 2011*. Institute of Education Sciences, U.S. Department of Education, Washington, D.C.
- Petersen, J. (2018). Gender difference in verbal performance: A meta-analysis of United States State performance assessments. *Educational Psychology Review*, 30(4), 1269–1281. <https://doi.org/10.1007/s10648-018-9450-x>
- Rauin, U. & Meier, U. (2007). Subjektive Einschätzungen des Kompetenzerwerbs in der Lehramtsausbildung [Subjective assessments of skills acquisition in teacher training]. In M. Lüders & J. Wissinger (Eds.), *Forschung zur Lehrerbildung, Kompetenzentwicklung und Programmevaluation* (pp. 102–131) Münster: Waxmann.
- Ready, D. D., & Wright, D. L. (2011). Accuracy and inaccuracy in teachers' perceptions of young children's cognitive abilities. *American Educational Research Journal*, 48(2), 335–360. <https://doi.org/10.3102/0002831210374874>
- Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender

- Ullman. (2020). The declarative/procedural model: A neurobiologically motivated theory of first and second language 1. In B. VanPatten, G. D. Keating & S. Wulff (Eds.), *Theories in Second Language Acquisition* (3<sup>rd</sup> ed., pp. 128–161). New York: Routledge.  
<https://doi.org/10.4324/9780429503986-7>
- Urhahne, D., & Wijnia, L. (2021). A review on the accuracy of teacher judgments. *Educational Research Review*, 32, 100374.  
<https://doi.org/10.1016/j.edurev.2020.100374>
- van Ewijk, R. (2011). Same work, lower grade? Student ethnicity and teachers' subjective assessments. *Economics of Education Review*, 30(5), 1045–1058.  
<https://doi.org/10.1016/j.econedurev.2011.05.008>
- Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgements of ESL argumentative essays. *Assessing Writing*, 39, 50–63.  
<https://doi.org/10.1016/j.asw.2018.12.003>
- Vögelin, C., Jansen, T., Keller, S. D., & Möller, J. (2018). The impact of vocabulary and spelling on judgments of ESL essays: an analysis of teacher comments. *The Language Learning Journal*, 49(6), 631–647.  
<https://doi.org/10.1080/09571736.2018.1522662>
- Weigle, S. C. (2007). Teaching writing teachers about assessment. *Journal of Second Language Writing*, 16(3), 194–209.  
<https://doi.org/10.1016/j.jslw.2007.07.004>

## Supplementary Material

### Figure S1

#### *Text assessment using the holistic scale in the Student Inventory*

**Text 1**

Eine Klage in Höhe von 67 Millionen Dollar

Ein Fall in der USA schockierte die Welt, da ein US-Richter, namens Roy Pearson, eine Reinigung auf Schadenersatz wegen Irreführung in Höhe von 67 Millionen Dollar verklagt.

Nachher reduzierte sich die Klage auf 54 Millionen Dollar, doch das Verwunderliche ist, dass es um eine abgegebene Hose handelt, die verloren gegangen ist. Die Klage wurde gestellt, da sich im Schaufenster Schilder mit dem Hinweis „Garantierte Zufriedenheit“ befanden und aufgrund des Wortes „Garantierte“ fordert der US-Richter ein Schadenersatz, weil die Garantie eine Entschädigung bereitstellen sollte. Die Reinigung hat am Ende die Schilder entfernt, doch ob sie nun aufgrund der Klage zahlen muss, steht noch offen.

Bitte beurteilen Sie auf dieser Seite den Text auf der linken Bildschirmseite. Nutzen Sie dazu die Skala auf der rechten Bildschirmseite.

Über dieses Optionsfeld können Sie sich erneut die Aufgabenstellung oder die inhaltlichen und stilistischen Erwartungen an den Text ansehen.

Ausblenden

Wählen Sie den Text aus, den Sie bewerten möchten. Sie müssen erst alle Texte beurteilt haben, bevor Sie "Weiter" klicken können.

Text 1

Text 1: noch nicht bearbeitet  
Text 2: noch nicht bearbeitet  
Text 3: noch nicht bearbeitet  
Text 4: noch nicht bearbeitet

**Globalskala**  
Sollten Sie die Beschreibungen der Skalenstufen erneut lesen wollen, klicken Sie auf die jeweilige Stufe.

☐ Stufe 5 - beste Bewertung  
☐ Stufe 4  
☐ Stufe 3  
☐ Stufe 2  
☐ Stufe 1 - schlechteste Bewertung

Weiter

### Expectation criteria and assessment scales (translated) of Study 1 and 2

#### Content expectations

For the newspaper article, all key points from the stimulus are used and processed analogously. Students are expected to prioritize the given information as follows (optional information is in brackets):

1. Beginning: main information

The name of the plaintiff and the exact amount of the claim are dispensable. However, the newspaper article must make it clear that the claim is reduced from the amount initially demanded or was originally even higher based on the last amount demanded. Functional additions: Information can be added to the aspects mentioned in the stimulus that serve to situate the event in time (e.g. "yesterday") or establish causality, for example:

- a reason for the reduction in demand is given (e.g. because of the removal of the signs)
- a reason is given for the removal of the signs (e.g. because of the high claim)

### **Stylistic expectations**

A newspaper report is to be written. The aim is to produce a coherent text, not a mere transcript of the stimulus. You are expected to use language that is appropriate to the addressee and relevant to the text type. Non-functional narrative and/or commentary parts are undesirable.

The appropriate tense for this type of text is used (simple past); the use of other tenses must be functional. Impertinent and colloquial formulations are to be devalued. An extensive vocabulary and varied sentence structure are to be considered positively. Elements appropriate to the type of text, such as headlines, are expected.

### **Global scale**

The student texts are to be assessed in relation to the writing occasion. Texts in the following categories have one or more of the following characteristics:

**Level 5** - The text develops and organizes the information with well-chosen details or the text presents all important given information accurately. The text has a successful structure and is coherent throughout. The type of text is adhered to throughout. The text is characterized by varied sentence structure and a mostly accurate choice of words. There are hardly any errors in grammar, spelling or comma placement and these do not impair the comprehension of the student text.

**Level 4** - The text develops and develops the information with some detail or the text presents almost all the important information given. The text has a clear structure, but may have a few flaws in coherence and/or occasional inconsistencies in causal or temporal sequence. The text type is largely adhered to. The text demonstrates confidence in sentence structure and in respecting sentence boundaries, but sentence structure and word choice are rather simple and lack variety. Errors in grammar, spelling or comma placement occur sporadically, but hardly affect the comprehension of the student's text.

**Level 3** - The text presents some information clearly, but appears to be listing, less developed or repetitive, or the text presents only some of the important information given. The text is irregularly structured, the statements occasionally seem disconnected. The text sometimes deviates from the appropriate text type. The text mostly shows certainty in sentence structure and in the observance of sentence boundaries; the choice of words is inappropriate or incorrect in some places. Errors in grammar, spelling or comma placement impair the comprehension of the student text in some places.

**Level 2** - The text presents information fragmentarily or is highly repetitive or poorly developed. The text lacks structure to a considerable extent, the statements are poorly connected or the development is too brief to allow a structure to be recognized. There is frequent deviation from the appropriate text type. The text shows little certainty in sentence structure and in the observance of sentence boundaries. The choice of words is often inappropriate or incorrect. Errors in grammar or in the use of language, in spelling and in the use of commas impair the comprehension of the student text in many parts.

**Level 1** - The text attempts to deal with the task, but provides little or no coherent information or merely paraphrases the task. The text shows no recognizable structure or consists of a single isolated statement. The text deviates from the appropriate text type. The text shows little or no command of sentence structure and little or no respect for sentence boundaries. The choice of words is predominantly incorrect or incorrect throughout the text. A variety of errors in grammar or language use, spelling and comma placement largely prevent comprehension of the entire student text.

### **Content scale**

The following criteria should be taken into account when evaluating the content:

- full consideration of the information provided
- consistency in the presentation of the information, taking into account its priority (main and additional information) and - if available - the conciseness and adequacy of the headline
- No introduction of own aspects and ideas that go beyond the given information and distort the context of the incident

**Level 4** - Level 4 texts take full account of the information provided. The presentation of the information is consistent and stringent in terms of content and follows the priority of the information given. No additional ideas of their own are included.

**Level 3** - Level 3 texts take full account of the information provided. The presentation of the information is mostly consistent in terms of content, but in some cases does not follow the priority of the information given, but rather the sequence of events. No further own ideas are included.

**Level 2** - Level 2 texts take into account the important information provided. The presentation of the information is largely coherent in terms of content, but does not follow the priority of the information given. Rarely do they include further ideas of their own that distort the context of the incident.

**Level 1** - Level 1 texts only take into account some of the information provided. The presentation of the information is not coherent in terms of content. Further own ideas are often included.

### **Style scale**

The following criteria should be taken into account when assessing the style:

- appropriate type of text (newspaper report)
- consistent perspective
- Creation of coherence / use of linking devices
- choice of words
- Variety and complexity of sentence construction
- Use of tense (past tense - simple past tense) Task-specific text elements:
- Headline

Sophisticated elements such as rhetorical devices and irony are to be considered favorably in the assessment.

**Level 4** - Level 4 texts adhere to the appropriate text type. The perspective is consistent and accurate. The texts appear coherent throughout due to the functional use of linking devices. The choice of words is always appropriate. The sentence structure is varied and not overly complex. The use of tense is consistent and appropriate. Task-specific text elements are present.

**Level 3** - Level 3 texts adhere to the appropriate text type. The perspective is consistent and appropriate. Linking devices are largely functional. Inconsistencies occur only occasionally. The choice of words is mostly appropriate. The sentence structure is mostly varied and not overly complex. The use of tense is mostly consistent and appropriate. Task-specific text elements are mostly present.

**Level 1** - Level 1 texts usually fall short of the appropriate text type. The perspective is sometimes unclear or inaccurate. Linking devices are not used or are not used functionally. Parts of the text appear incoherent. Inconsistencies occur frequently. The choice of words is hardly appropriate. The sentences are predominantly monotonous or incomprehensible. The use of tense is often inconsistent or inappropriate. Task-specific text elements are missing.

### **Linguistic accuracy scale**

The following features must be taken into account when assessing linguistic accuracy:

Correct spelling

Correct punctuation:

- Marking of sentence boundaries (sentence endings, capitalization at the beginning of sentences)
- Comma placement

Grammatical correctness, including with regard to

- Sentence structure
- Use of genus and case
- Inflection of words

**Level 4** - Level 4 texts meet the requirements for linguistic correctness according to the characteristics listed above. There are almost no errors in spelling, grammar and comma placement. The sentence boundaries are marked correctly throughout and all sentences are complete.

**Level 3** - Level 3 texts largely meet the requirements for linguistic correctness according to the characteristics listed above. There are hardly any errors in spelling, grammar and comma placement. The reading flow is not impaired. Very rarely are sentence boundaries incorrect or not marked or sentences incomplete.

**Level 2** - Level 2 texts partially meet the requirements for linguistic accuracy according to the characteristics listed above. There are errors in spelling, grammar and comma placement that impair the reading flow. In some cases, sentence boundaries are incorrect or not marked or sentences are incomplete.

**Level 1** - Level 1 texts hardly meet the requirements for linguistic accuracy according to the characteristics listed above. There are frequent grammatical and spelling errors that interrupt the reading flow. These include repeatedly ignored sentence boundaries or incomplete sentences. Commas are rarely used.

**Table S1**

*Results of the exploratory analysis regarding students' gender bias for teachers' assessment of medium-quality texts*

| Assessment scale    | Male name |           | Female name |           |                |          |                  |
|---------------------|-----------|-----------|-------------|-----------|----------------|----------|------------------|
|                     | <i>M</i>  | <i>SD</i> | <i>M</i>    | <i>SD</i> | <i>t</i> (116) | <i>p</i> | Cohen's <i>d</i> |
| Holistic            | 2.80      | 1.06      | 3.24        | 1.10      | 2.80           | .01      | .26              |
| Content             | 2.68      | 0.87      | 2.96        | 0.80      | 2.38           | .02      | .22              |
| Style               | 2.50      | 0.84      | 2.80        | 0.82      | 2.60           | .01      | .24              |
| Linguistic accuracy |           |           |             |           |                |          |                  |

**Table S2**

*Means and standard deviations of the four assessments for all three text quality levels of Study 1*

|                     | Text quality  |               |               |
|---------------------|---------------|---------------|---------------|
|                     | Low           | Medium        | High          |
| Scale               | <i>M (SD)</i> | <i>M (SD)</i> | <i>M (SD)</i> |
| Holistic            | 2.84 (1.23)   | 3.02 (1.10)   | 3.41 (0.92)   |
| Content             | 2.55 (0.85)   | 2.82 (0.85)   | 2.82 (0.85)   |
| Style               | 2.44 (0.96)   | 2.65 (0.84)   | 2.86 (0.77)   |
| Linguistic Accuracy | 2.59 (1.06)   | 2.65 (1.08)   | 3.10 (0.66)   |

*Note.* The holistic scale ranged from 1 to 5, and the analytic scales (content, style, linguistic accuracy) ranged from 1 to 4.

**Table S3**

*Post-hoc pairwise comparisons for text quality assessments of Study 1*

| Quality comparison  | <i>M<sub>diff</sub></i> | 95 % CI        | <i>t</i> (233) | <i>p</i> | <i>d</i> |
|---------------------|-------------------------|----------------|----------------|----------|----------|
| Holistic Assessment |                         |                |                |          |          |
| Low vs. medium      | -0.18                   | [-0.38, 0.02]  | -1.77          | .236     | -0.12    |
| Low vs. high        | -0.57                   | [-0.76, -0.38] | -6.01          | < .001   | -0.39    |
| Medium vs. high     | -0.39                   | [-0.56, -0.22] | -4.40          | < .001   | -0.29    |
| Content Assessment  |                         |                |                |          |          |
| Low vs. medium      | -0.27                   | [-             |                |          |          |

| Style Assessment    |       |                |       |        |       |
|---------------------|-------|----------------|-------|--------|-------|
| Low vs. medium      | -0.21 | [-0.37, -0.06] | -2.74 | .018   | -0.18 |
| Low vs. high        | -0.43 | [-0.58, -0.27] | -5.45 | < .001 | -0.36 |
| Medium vs. high     | -0.21 | [-0.35, 0.08]  | -3.05 | .008   | -0.20 |
| Linguistic accuracy |       |                |       |        |       |
| Low vs. medium      | -0.05 | [-0.24, 0.14]  | -0.53 | 1.00   | -0.03 |
| Low vs. high        | -0.50 | [-0.66, -0.35] | -6.31 | < .001 | -0.41 |
| Medium vs. high     | -0.45 | [-0.61, -0.30] | -5.88 | < .001 | -0.38 |

**Note.** Bonferroni-adjusted *p*-values were used for multiple comparisons.  $M_{diff}$  = Mean difference ( $M1 - M2$ ). CI = Confidence interval. Cohen's *d* was used to calculate effect sizes for dependent samples.

**Table S4**

*Means and standard deviations of the four assessments for all three text quality levels of Study 2*

|                     | Text quality  |               |               |
|---------------------|---------------|---------------|---------------|
|                     | Low           | Medium        | High          |
| Scale               | <i>M (SD)</i> | <i>M (SD)</i> | <i>M (SD)</i> |
| Holistic            | 2.73 (1.16)   | 2.96 (1.13)   | 3.34 (0.87)   |
| Content             | 2.44 (0.80)   | 2.72 (0.88)   | 2.79 (0.76)   |
| Style               | 2.33 (0.90)   | 2.67 (0.80)   | 2.84 (0.76)   |
| Linguistic Accuracy | 2.52 (1.11)   | 2.65 (1.04)   | 3.0           |

**Table S5**

*Post-hoc pairwise comparisons for text quality assessments of Study 2*

| Quality comparison  | $M_{diff}$ | 95 % CI        | $t(233)$ | $p$    | $d$   |
|---------------------|------------|----------------|----------|--------|-------|
| Holistic Assessment |            |                |          |        |       |
| Low vs. medium      | -0.23      | [-0.42, 0.04]  | -2.34    | .060   | -0.15 |
| Low vs. high        | -0.61      | [-0.77, -0.46] | -7.94    | < .001 | -0.50 |
| Medium vs. high     | -0.39      | [-0.55, -0.23] | -4.74    | < .001 | -0.30 |
| Content Assessment  |            |                |          |        |       |
| Low vs. medium      | -0.28      | [-0.41, -0.15] | -4.23    | < .001 | -0.27 |
| Low vs. high        | -0.35      | [-0.48, -0.23] | -5.55    | < .001 | -0.35 |
| Medium vs. high     | -0.07      | [-0.20, 0.05]  | -1.15    | .753   | -0.07 |
| Style Assessment    |            |                |          |        |       |
| Low vs. medium      | -0.34      | [-0.49, -0.19] | -4.47    | < .001 | -0.28 |
| Low vs. high        | -0.51      | [-0.64, -0.37] | -7.46    | < .001 | -0.47 |
| Medium vs. high     | -0.17      | [-0.30, -0.03] | -2.48    | .041   | -0.16 |
| Linguistic accuracy |            |                |          |        |       |
| Low vs. medium      | -0.13      | [-0.32, 0.06]  | -1.37    | .513   | -0.09 |