Original Article

Feedback on teachers' text assessment: Does it foster assessment accuracy and motivation?

Thorben Jansen[1,2], Jennifer Meyer[1], Stefan Schipolowski[3], and Jens Möller[2]

[1]IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany
[2]Institute for Psychology of Learning and Instruction (IPL), Kiel University, Germany
[3]Institute for Educational Quality Improvement (IQB), Humboldt-Universität zu Berlin, Germany

Abstract: Teachers' assessment of students' performance on complex tasks, such as writing, is important both for their teaching and for students' learning. Teachers must be able and motivated to assess texts correctly. According to theoretical assumptions, feedback can help promote the diagnostic competencies required to assess texts correctly, but, up until now, no empirical studies have examined the effects of accuracy feedback on teachers' assessments. We conducted an experimental study comparing the effects of two feedback interventions with a practice-only control group on teachers' assessment accuracy and motivation. Student teachers (n = 181) and experienced teachers (n = 114) assessed 10 students' texts in all groups. The feedback in both of the feedback groups showed the teachers a comparison between their own assessments and correct assessments. We varied the feedback presentation between one single presentation after five texts and single presentations after each of the first five texts. We measured assessment accuracy and situational interest, which conceptualizes motivation, to assess the next five texts. The results showed that feedback promoted situational interest but not assessment accuracy. We discuss why teachers found feedback interesting and under what circumstances training interventions could be useful.

Keywords: diagnostic competencies, assessment accuracy, interest, motivation, training

Erhöht Feedback zur Qualität ihrer Urteile die Beurteilungsqualität und Motivation von Lehrkräften?

Zusammenfassung: Die Beurteilungen von Lehrkräften zu den schriftlichen Leistungen ihrer Schülerinnen und Schüler spielen eine wichtige Rolle für die Unterrichtsgestaltung und das Lernen. Daher sollen für die Aus- und Weiterbildung von Lehrkräften Trainingsmöglichkeiten geschaffen werden, welche die Kompetenz und Motivation der Lehrkräfte zur korrekten Beurteilung erhöhen. Es ist anzunehmen, dass Feedback diagnostische Kompetenzen fördern kann, aber momentan fehlt es an empirischen Studien, die die Effektivität von Feedback im Vergleich zu Beurteilungsübungen ohne Feedback untersuchen. Der vorliegende Artikel stellt eine experimentelle Studie vor, in der die Effekte von zwei Feedback-Interventionen auf die Beurteilungsgenauigkeit und Motivation der Lehrkräfte mit einer Kontrollgruppe, in der Texte ohne Feedback beurteilt wurden, verglichen wurden. Lehramtsstudierende (n = 181) und erfahrene Lehrkräfte (n = 114) bewerteten in allen Gruppen zehn Schülertexte. In beiden Feedbackbedingungen wurde den Lehrkräften ein Vergleich der eigenen Bewertung mit der Bewertung des Textes durch Expert:innen gezeigt. Zwischen den Gruppen variierte die Feedback-Präsentation zwischen einer einmaligen Präsentation nach fünf Texten und mehrmaliger Präsentation nach jedem der ersten fünf Texte. Danach wurde die Motivation weitere Texte zu beurteilen, sowie die Beurteilungsgenauigkeit bei der Beurteilung der zweiten fünf Texte gemessen. Die Ergebnisse zeigten, dass beide Feedbackbedingungen im Vergleich zur Kontrollgruppe das situative Interesse an der Beurteilung förderten, aber nicht die Genauigkeit. Es wird diskutiert, warum Lehrkräfte die Beurteilung mit Feedback interessanter fanden und unter welchen Umständen Trainingsinterventionen nützlich sein können.

Schlüsselwörter: Diagnosekompetenzen, Beurteilungsgenauigkeit, Interesse, Motivation, Ausbildung

Highlights

- Experimental study that allowed teachers to compare their assessment with expert judgments
- Text assessment was investigated on global and analytical scales
- We found evidence of positive effects of feedback on situational interest in text assessment
- No significant differences were found between the feedback and control conditions for assessment accuracy

Teachers' diagnostic competence is an essential component of their professional competencies; it describes teachers' ability and motivation to assess students' performance accurately (Baumert & Kunter, 2006). Teachers' assessments and the resulting judgments function as a prere-

https://doi.org/10.1024/1010-0652/a000365

quisite to adapting lessons to students' competencies (Herppich et al., 2017), which form the basis for students' self-concepts (Möller et al., 2020), and they serve as performance feedback for students (Elliott et al., 2001). The empirical literature on teachers' diagnostic competence has either focused on the assessment of specific aspects of student performance such as their writing (teacher assessments; e.g., Jansen et al., 2021; Möller et al., 2022) or examined the accuracy of teachers' knowledge of student characteristics on the basis of completed assessments (teacher judgments; see Südkamp et al., 2012; Urhahne & Wijnia, 2021, for an overview). In this article, we use the term "teachers' assessments" to refer to teachers' assessments of specific aspects of students' performance, for example, teachers' grading of students' written exams after a class test. In these contexts, assessments are important because they allow teachers to provide students with the most effective feedback and subsequent learning opportunities, for example, by providing additional instruction or adapted tasks.

Teachers have difficulties in accurately assessing students' writing (Birkel & Birkel, 2002; Möller et al., 2022; Jansen et al., 2021b). Texts represent multifaceted performances (Sadler, 1989), in which the assessment of one facet may bias the assessment of other facets, as shown for text length (Fleckenstein et al., 2020; Rezaei & Lovorn, 2010), organization (Crossley et al., 2014; Vögelin et al., 2020), or the vocabulary of students' texts (Scannell & Marshall, 1966; Vögelin et al., 2019). Overcoming these difficulties and raising teachers' assessment accuracy is an essential aim of teacher education (Chernikova et al., 2020).

In addition to fostering assessment accuracy, an aim of assessment training, in particular, and teacher education, in general, is to increase teachers' assessment motivation because teachers' motivational conditions are considered to be a part (Baumert & Kunter, 2006; Weinert, 2001) or a predictor (Herppich et al., 2019) of teachers' diagnostic competence. Empirical evidence has shown that teachers' motivation has a predictive effect on their diagnostic competence (Klug et al., 2016; Kron et al., 2022). Especially during the long and cognitively demanding assessment of a whole class of students' writing performance, teachers need to stay motivated to make a fair assessment and to give effective feedback. Correct and meaningful feedback is an important instructional tool that can help foster students' writing competencies.

Several interventions have had the goal of promoting diagnostic competencies (see Chernikova et al., 2019, for an overview) by increasing assessment quality (Baird et al., 2017; Jansen et al., 2021a) and motivation (DeLuca et al., 2013; Dempsey et al., 2009). Previous training programs on the assessment of students' written work were conducted over several days to prepare teachers to assess large-scale writing studies (Chamberlain & Taylor, 2011; Choi & Wolfe, 2020; Raczynski et al., 2015). They included the assessment of example texts, receiving accuracy feedback that compared teachers' assessments to expert judgments, discussions with expert raters, input videos, and prompts.

However, these training programs combined multiple interventions to achieve the largest effects and did not evaluate the effects of individual interventions such as accuracy feedback compared to a practice-only control group. As a result, it remains unclear which interventions caused the positive effects. Investigating the effects of individual interventions is of particular interest because shorter interventions are more accessible for teachers than multiday training and they have been shown to have positive effects on teachers' professional competencies (Heitzmann et al., 2018; Ohst et al., 2015; Reeves & Chiang, 2017). Moreover, such short interventions are an established part of teacher training (see Merk et al., 2023 for an evaluation of one exemplary intervention). In our study, we tested the effects of two types of feedback intervention on assessment accuracy and motivation in an experimental study.

Performance feedback in the domain of assessment accuracy: Theoretical assumptions

Theoretical models (Heitzmann et al., 2019) and meta-analyses (Chernikova et al., 2019) on the fostering of teachers' diagnostic competencies describe interventions, including feedback, as an effective method to promote teachers' diagnostic competencies. Established interventions that aim to foster diagnostic competencies include feedback (e.g., Chamberlain & Taylor, 2011; Choi & Wolfe, 2020; Dempsey et al., 2009; Raczynski et al., 2015). Feedback can be an effective instructional practice because it enables teachers to close the gap between their performance (i.e., their assessment) and the desired target (i.e., an accurate assessment; Biber et al., 2011; Hattie & Timperley, 2007). With that aim, feedback should specify learning targets, include evaluations of the current state of the learning process, and give information about the next step toward the targets (Black & William, 2009; Shute, 2008). The cognitive process of receiving feedback is described in the Interactive-Two-Feedback-Loops-Model (Narciss, 2007). The model differentiates between two interacting feedback loops: the feedback receiver loop and the feedback source loop. The internal receiver loop contains the teachers' (the feedback receiver in this case) representations of competencies and the related standards, which serve as a basis for the internal reference value regarding the targeted competence standards. The external feedback source loop includes an external representation

of standards, competencies, and task requirements and serves as a basis for determining external reference values and standards. When receiving external feedback, for example, in the form of expert judgments, the receiver compares the internal and external standards within the receiver loop. If the comparison shows a gap between both standards, the feedback receiver can select actions to close the gap.

In our study, the theoretical mechanisms through which we assume feedback interventions to work are based on the Interactive-Two-Feedback-Loops-Model (Narciss, 2007). To improve assessment performance, we designed our feedback to inform the receiver about the gap between the actual and the target values, thereby giving them the opportunity to restructure their internal standards. In this way, the intervention aimed to make the teachers aware of how their assessments deviated from the desired standards, which could help them to adjust and refine their internal standards for assessment. Also, our feedback aimed to support the teachers by showing the direction in which their assessments deviated from the desired standards; the feedback provided was specific enough to enable teachers to act on it during the next assessments, so that these next assessments could align better with the target standards.

Feedback can also have motivational benefits, for example, by increasing teachers' self-efficacy. In our study, the feedback provided after the first five texts may have reduced the (perceived) assessment difficulty by showing the teachers their high performance on some scales in the assessment of the first five texts. If teachers see what they have already assessed correctly, this can result in a sense of competence, which can lead to higher interest in assessing the next texts and in performing well (Vu et al., 2022). For teachers who did not perform well on the first five texts, the feedback showed them how to improve on the scales on which they had not made accurate assessments. By knowing how they can improve in the following task, they also might experience higher self-efficacy, and this might lead to more interest in the task (Spinath & Steinmayr, 2012).

Training diagnostic competence in assessing writing: Empirical studies

Meta-analytic evidence suggests that interventions can increase teachers' diagnostic competence (Chernikova et al., 2019). For example, prompt interventions and reflection phases showed positive mean effects in medium sizes (Hedges' g = 0.47 and 0.58). The meta-analysis included accuracy and motivation measures as outcomes but did not differentiate between them. The following section summarizes the results of empirical studies on interventions that have been conducted with the aim of increasing the accuracy of teachers' assessments or motivation when assessing students' written performance.

Fostering assessment accuracy

Baird and colleagues (2017) showed positive assessment training effects within the assessment of England's 2008 national curriculum English writing test for 14-year-olds (General Certificate of Secondary Education – GCSE) by comparing the assessment accuracy of teachers who had participated in several years of assessment training with that of teachers who were assessing the writing test for the first time. The training included multiple training days, during which the participants received feedback on how their assessments compared with expert judgments; they also assessed texts in rating groups, participated in discussions, and listened to input talks from team leaders. However, the study did not contain a practice-only control group (Royal-Dawson & Baird, 2009; Leckie & Baird, 2011). Chamberlain and Taylor (2011) compared the GSCE online and face-to-face training programs and found similar positive effects on assessment accuracy in a pre-post training comparison. In another assessment training program for a large-scale writing study, Raczynski and colleagues (2015) investigated whether feedback that compared teachers' assessments with expert judgments should be better provided one-on-one with the expert or in a group. Both types of training showed positive effects, which did not differ in strength. Choi and Wolfe (2020) trained participants over several days with feedback that compared their assessment with expert judgments. They varied whether or not the participants additionally discussed the feedback with a mentor. Their results showed positive effects for both groups, but accuracy increased more when participants talked to the mentor. Rethinasamy (2021) varied whether the participants read through text examples including expert judgments (control condition), evaluated text examples and compared their assessment with the expert judgments (Condition A), or read through text examples that included expert judgments, then evaluated other text examples, and finally compared their assessments with the expert judgments (Condition B). In comparison with participants in the control condition, participants in the training conditions showed higher assessment accuracy immediately after the training (Condition A) and in a delayed assessment test (Conditions A and B).

This brief overview of studies on how to foster assessment accuracy shows that participants in all of the studies received feedback that enabled them to compare their assessments with expert judgments and to repeat this comparison frequently over several days. In addition, the training combined this feedback with discussions or expert talks. Such training is well suited for improving experts'

assessments but not for fostering teachers' diagnostic competencies because of the great time and effort required of the teachers. For teachers, it would be desirable to have an effective self-learning module on feedback, which does not require discussion with others. Moreover, up until now, no study has examined the effects of feedback in comparison to a practice-only control group.

Fostering assessment motivation

Only three studies so far have investigated the motivational outcomes of diagnostic competence training. Dempsey and colleagues (2009) showed that participants, mostly psychology students, who completed several rounds of assessing a text on six scales (e.g., ideas, conventions) and who received feedback were more confident in their assessment than before the feedback. Feedback included prompts by a virtual coach, graphical peer feedback, verbal peer feedback, peer interaction, and expert feedback. DeLuca and colleagues (2013) also found positive effects of a semester-long assessment course on the assessment confidence of preservice teachers. Sommerhoff and colleagues (2022) examined the effects of providing preservice teachers with prompts on how to assess students' performance within a simulation and found no effect on teachers' assessment motivation.

In summary, all diagnostic competence training in writing assessment involved a comparison of participants' assessments with experts' judgments as a feedback measure that was combined with discussions with peers or experts. The results show, on average, that training has positive effects on assessment accuracy and motivation across all domains (Chernikova et al., 2019) and in assessing writing (e.g., Baird et al., 2017, Dempsey et al., 2009). However, no study thus far has included a practice-only control group; therefore, the feedback effects cannot be disentangled from the practice effects. Further, all studies repeated the feedback many times because it was part of a multiday training intervention in the context of a large-scale writing assessment; this means that it is not easily accessible for teachers.

The present study

Student teachers' and experienced teachers' accuracy in assessing writing performance needs to be improved (e.g., Birkel & Birkel; Jansen et al., 2021b). Established assessment training programs are often very extensive and time-consuming, making participation unattractive for teachers. To promote teachers' diagnostic competence, it is important to develop simple and easy-to-implement training programs. With this aim, this study investigated the effects of accuracy feedback as a training component. In the study, we compared two types of feedback intervention with the aim of improving not only the agreement between teachers' assessments and experts' judgments but also teachers' assessment motivation, which we measured as situational interest. Our study included a multiple feedback condition, as in previous studies, and a more economic single feedback condition. In the multiple feedback condition, feedback was shown immediately after teachers had assessed each text. In the single feedback condition, feedback showed teachers their average assessment accuracy over the first five texts. This study extends the literature twofold: first, by experimentally investigating the effectiveness of both a single and a multiple feedback intervention in the domain of writing assessment compared to a practice-only control group and, second, by including situational interest as a motivational variable in addition to assessment accuracy as a dependent variable. We tested the following two hypotheses:

Hypotheses on teachers' assessment accuracy:

H1. We expected to find differences in the percentage of accurate assessments between the single feedback group, the multiple feedback group, and the practice-only group.

H1a. We expected the multiple feedback group to have stronger positive effects on the percentage of accurate assessments than the practice-only group.

H1b. We expected the single feedback group to have stronger positive effects on the percentage of accurate assessments than the practice-only group.

Hypotheses on teachers' situational interest:

H2. We expected to find differences in the situational interest between the single feedback group, the multiple feedback group, and the practice-only group.

H2a. We expected the multiple feedback group to have stronger positive effects on the situational interest to assess texts than the practice-only group.

H2b. We expected the single feedback group to have stronger positive effects on the situational interest to assess texts than the practice-only group.

We tested the difference between the multiple and single feedback groups exploratively for both outcomes as we assumed that the two groups had the same psychological mechanism but we assumed that the feedback design would fulfill the cognitive and motivational feedback functions to different degrees. The two conditions differed in their presentation of the gap between teachers' assessments and experts' judgments. On the one hand, in the single feedback condition, comparing one's own assessment to expert judgments was easier because teachers could directly see their tendency to assess too strictly or too leniently. However, the single aggregated score did not provide information about the individual text qualities, making it more difficult to draw conclusions about what exactly could be improved in the text assessment. On the other hand, the multiple feedback condition made it possible to give feedback on specific

texts, supporting teachers to create specific hypotheses about their assessment standards in relation to the expert assessments. Accordingly, on the basis of the feedback, teachers could adjust their assessment and then receive the next feedback to check again. However, the process of multiply evaluating one's own assessment in comparison to that of experts can be cognitively demanding, which can lead to cognitive overload, resulting in a reduced effectiveness of the multiple feedback.

Method

Transparency and openness

In the following sections, we report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study. We followed the journal article reporting standards (e.g., JARS; Kazak, 2018). All data, analysis code, and research materials are available at https://osf.io/mq8pk/. Data were analyzed using SPSS Version 28. This study's design and its analysis were not preregistered. The study was reviewed by the ethics committee of the IPN-Leibniz Institute for Science and Mathematics Education in Kiel.

Sample

We calculated the target sample size using G*Power (Faul et al., 2007) to find a standardized difference of 0.40 between groups with a power of .80, controlling for two covariates in a fixed-effects analysis of variance. We wanted to test this medium standardized mean difference so that our feedback would increase diagnostic competence similar to an average intervention in higher education (Chernikova et al., 2019). The target sample size was $N = 244$.

The sample consisted of $N = 295$ teachers, including both student teachers ($n = 181$) and experienced teachers ($n = 114$). The sample of student teachers was acquired through advertisements within regular university seminars, whose curriculum did not include assessments of students' performance. The study took place outside the seminar hours and only a few people from each seminar participated. The student teachers had, on average, been enrolled in the teacher education program for 8.71 semesters ($SD = 8.50$). Of the student teachers, 120 identified their gender as female, 56 as male, and two as nonbinary.

Experienced teachers participated in the study within a professional development program or responded to our advertisements in schools. The experienced teachers in our sample had a mean teaching experience of 10.94 years ($SD = 11.55$), starting with the completion of their master's degree. Of the experienced teachers, 80 identified their gender as female, 33 as male, and one as nonbinary.

The sample was randomly divided into the no-feedback (control) group (experienced teachers: $n = 31$, student teachers: $n = 54$), the single feedback group (experienced teachers: $n = 37$, student teachers: $n = 61$), and the multiple feedback group (experienced teachers: $n = 56$, student teachers: $n = 66$). The groups did not significantly differ in the number of student teachers and experienced teachers (see ESM 1, Table S1). We did not exclude any participants from our analyses.

Materials

Selection of texts

Participants assessed a text sample of the same 10 students' texts and compared their assessments to expert judgments within the first five texts. We randomized the order of the 10 texts between the participants. In selecting the text tasks, the texts, and the associated expert judgments, we drew on previously proven work from a norming study for a large-scale assessment study representative for Germany (Canz, 2015; 2021). The text corpus included 300 texts from eighth-grade students in the German language. We selected the 10 texts with the aim of them being representative of the corpus in terms of the expert judgments and the text length (see ESM 1, Figure S1 for the empirical distribution).

The experts who produced the ratings completed an intense training phase, with three sets of trial ratings for the writing task, as well as three whole-day training days. Within each trial rating, the experts assessed a package of 20–40 student texts and discussed their assessment. On the training days, experts were trained to ignore unintended aspects, such as orthographic and grammatical errors, in the content and style scoring. This intense training was implemented to guarantee the interchangeability of raters and, consequently, the reliability of the scores. For each text within the study, the judgments of the two experts were in agreement with each other.

Assessment task

The assessment task was to assess texts produced by students the mentioned writing assignment. The writing assignment was to write a newspaper article about a judge suing a dry cleaner over a pair of damaged pants. A team of linguists, psychologists, psychometricians, and teachers of German developed the prompts and assessment scales based on the assessment scales of the National Assessment of Educational Progress (NAEP; National Assessment Governing Board, 2011a, 2011b; National Center for

Education Statistics, 2012). The texts were handwritten and about one page long. The texts' structure was mostly a chronological report or not recognizable.

Assessment scales

The assessment scales included a global scale and the three analytical scales, content, style, and mechanics. The raters were extensively trained for the texts and rating scales (see Canz, 2015, for the training procedure and assessment scales). The interrater reliability, calculated as the percentage of exact agreement and the intraclass correlation coefficient (ICC), on the five-point global scale (exact agreement = 52%; ICC = .64) and on the four-point scales for content (exact agreement = 60%; ICC = .66), style (exact agreement = 57%; ICC = .55), and mechanics (exact agreement = 59%; ICC = .67) were similar to previously reported reliabilities for writing assessments. Brown and colleagues (2004) reported exact rater agreement ranging between 40% and 60%. The ICCs were "moderate" (Koo & Li, 2016) and comparable to those found in other large-scale writing assessments (e.g., Keller et al., 2020) and to teachers' reliability in assessing writing after a long training program (Skar & Jolle, 2017).

Independent variable: Feedback

The feedback was randomly varied between participants into three levels. The feedback was shown within or after the first five texts. The feedback levels were the control group (no feedback), the single feedback group (single feedback given after the first five texts, comparing participants' assessments with expert judgments), and multiple feedback (feedback given after each of the first five texts, comparing participants' assessments with expert judgments). Each feedback showed a bar chart with four bars for the participants' assessments on the four rating scales (i.e., global, content, style, and mechanics) and four bars for the experts' judgments. Participants were informed that the bars of their assessments and those of the experts' judgments should have the same height (ESM 1, see the left part of Figure 1 for the multiple feedback condition and Figure S2 for the single feedback condition; the figures were presented to the participants in German). Please note that the participants saw only the left part of Figure 1. On the right side of the figure, we have added descriptions that link the feedback components to the cognitive and motivational feedback functions described in the interactive two-feedback-loops model (Narciss, 2007).

Dependent variables

Assessment accuracy after feedback

We measured the accuracy of the assessment of Texts 6–10, which were identical for each group. We calculated the mean difference between teachers' assessments and experts' judgments (Südkamp et al., 2008) to show strict

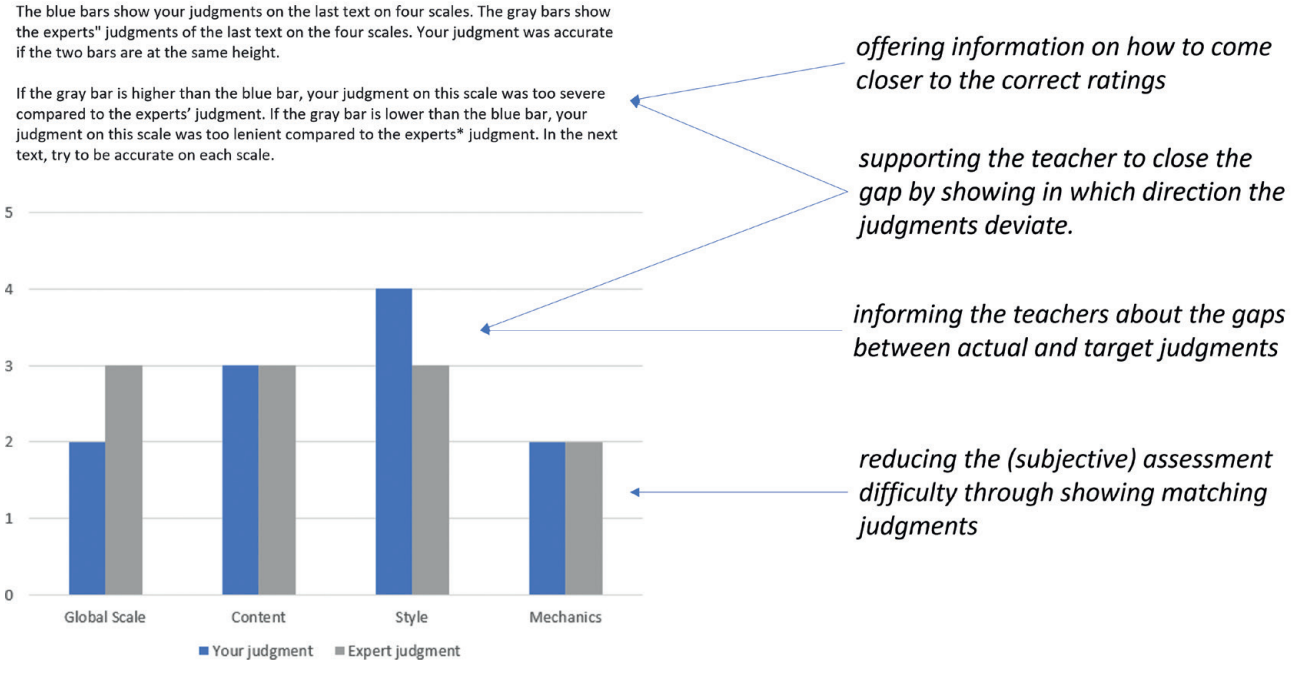

Figure 1. Feedback in the Multiple Feedback Condition (left side) With Assumed Cognitive and Motivational Function (right side).

and lenient tendencies averaged across texts. Additionally, we calculated the hit rate as the percentage of exact agreement between teachers' assessments and experts' judgments to show the accuracy of each text.

Situational interest

We operationalized assessment motivation as situational interest because it is a prerequisite for the development of long-term interest and is easily triggered by interventions (Hidi & Renninger, 2006). We measured situational interest with six items from Rotgans and colleagues (2014), adapted to the context of assessing texts. After assessing five texts and seeing the feedback, participants assessed how much they agreed with six items on a six-point scale, ranging from *completely agree* to *completely disagree*. The items are: "I enjoy working on this task", "I want to know more about this task", "I think this task is interesting", "I expect to master this task well", "I am fully focused on this task; I am not distracted by other things," and "Presently, I feel bored" (reversed). The reliability of the situational interest scale was .82.

Covariates

Theoretical models (Heitzmann et al., 2019; Herppich et al., 2018) include teachers' experience and qualifications as moderators of assessment accuracy, and some studies have provided evidence for their relation to assessment accuracy (e.g., Jansen et al., 2021b; McElvany et al., 2009; Möller et al., 2022). Therefore, as a robustness check, we tested whether our data supported the hypotheses by also controlling for teachers' experience and qualifications.

Teachers' experience

We divided this covariate into two levels (experienced teachers vs. student teachers). Of the 295 participants, 181 were student teachers (mean age = 25.83 years [*SD* = 7.40], 66 % female) and 114 were experienced teachers (mean age = 38.66 years [*SD* = 13.07], 70 % female).

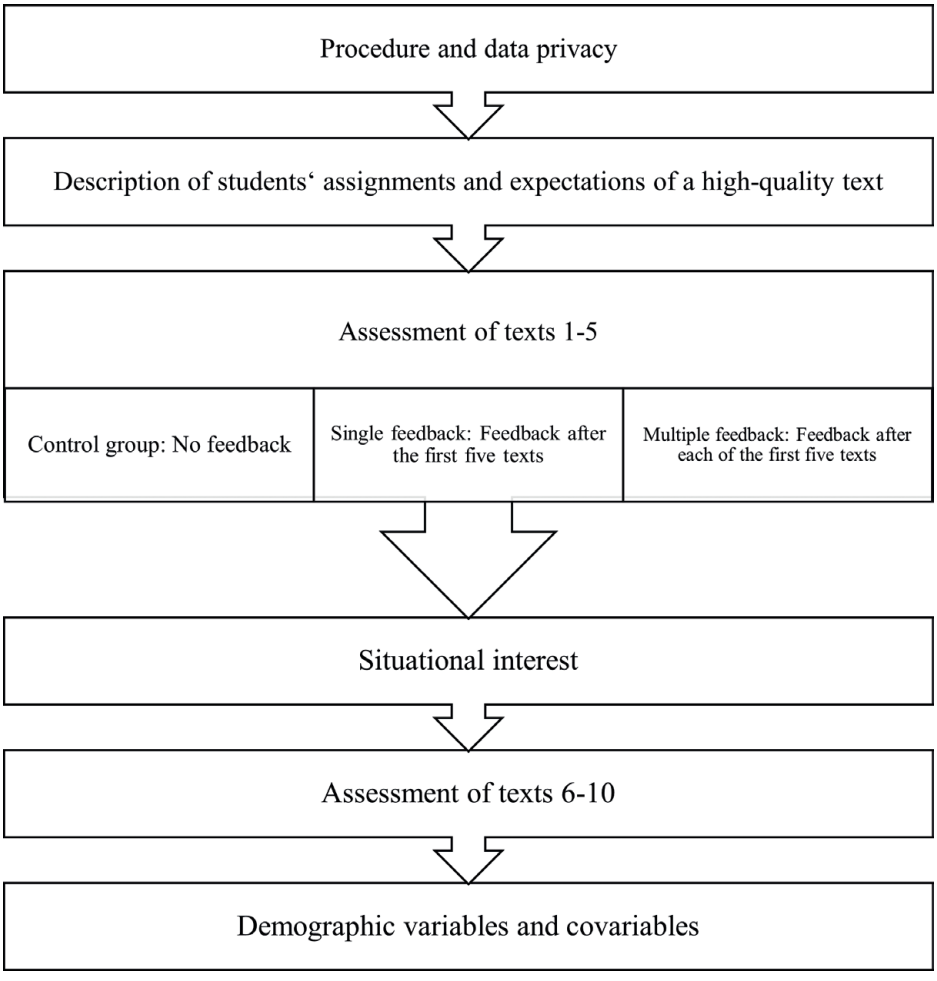

Figure 2. Study procedure

Table 1. Means (Standard deviations) for assessment accuracy, split by group and scale

	Mean difference				Hit rate			
	Global	Content	Style	Mechanics	Global	Content	Style	Mechanics
No feedback	−0.01 (0.64)	−0.07 (0.48)	0.48 (0.51)	0.13 (0.44)	0.38 (0.53)	0.40 (0.53)	0.32 (0.51)	0.42 (0.54)
Single feedback	−0.04 (0.54)	−0.12 (0.47)	0.41 (0.57)	0.06 (0.47)	0.38 (0.49)	0.42 (0.50)	0.32 (0.47)	0.39 (0.49)
Multiple feedback	−0.08 (0.62)	−0.11 (0.50)	0.44 (0.54)	0.16 (0.49)	0.31 (0.44)	0.39 (0.46)	0.33 (0.44)	0.39 (0.46)

Note: Mean difference is the averaged difference across the last five texts between teachers' assessments and experts' judgments. Correct assessments would result in a value of zero. The hit rate is the relative frequency of exact agreement between teachers' and experts' judgments. Correct assessments would result in a value of one.

Major in German studies

We divided this covariate into two levels (teachers with a major in German studies and teachers without a major in German studies). The sample contained 132 participants who majored in German studies (mean age = 29.23 years [*SD* = 11.17], 77% female) and 161 who did not major in German studies (mean age = 32.16 years [*SD* = 12.16], 61% female). Two participants did not report if they majored in German studies.

Procedure

Participants assessed 10 student texts on a global and on three analytic assessment scales (content, style, and mechanics). Each participant assessed the same texts in a randomized order. We conducted the study in the online assessment tool "Schülerinventar" (The student inventory; Jansen et al., 2019). The tool has already been used for studies on teachers' assessments in subjects such as English as a foreign language (Jansen et al.; 2021a; Vögelin et al., 2018; 2019), biology (Fischer et al., 2021), math and history (Jansen et al., 2022). The online tool is a sequence of web pages that participants access one after the other at their own speed. On average, participants took 36 minutes to complete the study. Before starting the study, the participants had to sign a confidentiality declaration about the texts and the scales shown. With this declaration, participants confirmed that they would not distribute the materials as the materials will be used in future school assessments. The participants saw the same materials and rating scales as the experts, including highly detailed rubrics for rating text quality and examples. The participants did not receive money or any other incentives for their participation. Figure 2 shows the study procedure.

Statistical analyses of diagnostic competence

Teachers' assessment accuracy is a key indicator of teachers' performance in a diagnostic situation and measures how well the teacher's assessment matches the student's performance (Jansen et al., 2021b). A measure of accuracy for single texts is the percentage of agreement between teachers' assessments and experts' judgments (e.g., Baird et al., 2017; Raczynski et al., 2015). A measure of the average accuracy across multiple texts is the component level, that is, the difference between teachers' and experts' judgments averaged across texts (Jansen et al., 2021b). We compared the accuracy components between groups using univariate analyses of variance for each component and each scale with the feedback groups as a three-level factor. We further analyzed significant main effects by conducting pairwise *t*-test comparisons between groups.

Results

The randomization check showed no significant differences in assessment accuracy in the first five texts on the global and the analytical scales between the groups (see ESM 1: Table S1, and Figures S4 and S5). All three groups had a similar percentage of experienced teachers, teachers with a major in German studies, and female teachers (see ESM 1, Table S2). Therefore, we consider the randomization to have been successful.

Assessment accuracy

Table 1 shows the descriptive statistics for the accuracy of the assessments of the last five texts, split by the scale. Regarding the mean differences averaged across the three groups, participants' assessments either did not differ significantly from the perfect value on the global scale ($t[294] = 1.30$, $p = .195$, $d = 0.08$), were too lenient on the content scale ($t[294] = 3.62$, $p < .001$, $d = 0.21$), or were too strict on the style ($t[294] = -14.00$, $p < .001$, $d = 0.82$) and mechanics ($t[294] = -4.26$, $p < .001$, $d = 0.25$) scales. Regarding the hit rates, participants' hit rates were lower than the hit rates of the experts on the global scale ($t[294] = -12.07$, $p < .001$,

d = 0.70) and the analytical scales of content (t[294] = –14.66, p < .001, d = 0.85), style (t[294] = –19.87, p < .001, d = 1.16), and mechanics (t[294] = –16.13, p < .001, d = 0.94). We tested whether the assessment accuracy, averaged across all groups, improved from the assessment of the first five to the last five texts (see ESM 1, also Figure S6). Our analyses showed no significant difference for the four scales or for the two accuracy measures (max t[294] = 1.66, p > .097 for the hit rate on the style scale).

We did not find any systematic significant differences between the feedback conditions in the analyses with or without the covariates (see ESM 1, Table S3 for the test statistics). As a robustness check, we also tested the results for other measures of assessment accuracy (quadratic-weighted kappa, correlation between participants' assessments and experts' judgments), and no differences were found there either (see ESM 1, Figure S4). One exception was a significant difference between the groups for the hit rate on the global scale (F[2, 292] = 3.51, p < .05). The pairwise comparison showed a significantly lower hit rate for the multiple feedback group compared to the single feedback group (d = –0.14; p = .024) and the control group (d = –0.15; p = .026). We did not find any difference between the single feedback and the no-feedback control groups (d = 0.01; p = .996).

Situational interest

We compared the situational interest components between groups using univariate analyses with the feedback groups as a three-level factor. Regarding the situational interest hypotheses, we found significant differences between groups (F[2, 292] = 3.90, p = .021), even when controlling for the covariates of teachers' experience and qualifications (F[2, 292] = 4.43, p = .013). Pairwise comparisons showed that the participants in the single feedback (M = 3.69, SD = 0.20) and the multiple feedback (M = 3.71, SD = 0.20) groups showed more situational interest (d = 0.71, p = .016, d = 0.71, p < .013) in assessing texts after the feedback than the control group did (M = 3.42, SD = 0.18). There was no difference between the single and the multiple feedback groups (d = 0.05, p < .885).

Discussion

We investigated a feedback intervention that compared teachers' assessments with experts' judgments with the aim of fostering teachers' diagnostic competence and motivation. We hypothesized that the feedback would reduce teachers' inaccuracies in their assessment of students' performance because seeing expert judgments would restructure teachers' internal standards by informing the teachers about the gap between their assessments and expert judgments, thereby supporting them to close the gap without immediately offering the solution (Narciss, 2018). Our findings did not support this hypothesis: We found no evidence of significant differences between the feedback conditions and the control group. One reason for this result might be that the mean difference between teachers' assessments and expert judgments before the intervention (see ESM 1, Table S1) was closer to zero than it was in previous studies (e.g., Baird et al., 2017; Chamberlain & Taylor, 2011; Jansen et al., 2021b; Royal-Dawson & Baird, 2009). Maybe the very detailed materials and assessment scales provided enough information to help even the inexperienced teachers to make correct assessments regarding students' text quality. Due to this high assessment accuracy, when teachers compared their own assessments with the expert judgments, a gap that might have motivated teachers to change their assessment standards was not obvious; this, in turn, might be the reason for why the feedback showed no effect. This fits with the idea that teachers only consider data to be meaningful for their actions when they show moderate to large deficits (Merk et al., 2023). Although we did not find a significant difference between teachers' assessments and expert judgments on the global scale, assessments on the more specific scales (content, style, and mechanics) and accuracy measures (hit rate, rank component, and quadratic weighted kappa) showed much room for improvement. The differences in content, style, and mechanics suggest that these assessments could lead to nonoptimal decisions in the creation of feedback or adaptive instruction. For example, an overly strict assessment on the style scale together with an overly lenient assessment on the content scale could lead teachers to a nonoptimal instructional focus on the promotion of style, even though instruction on the content of the text would be better in order to foster student learning. We consider the inconsistent pattern between the scales to be an indication of a strong dependence of assessment accuracy on the distribution of expert judgments, the scales, and the measures of accuracy. This interpretation fits with the literature on teacher judgment accuracy, which has meta-analytically linked judgment characteristics to teachers' accuracy (Südkamp et al. 2012) and showed low correlations between the components of judgment accuracy (Spinath et al., 2005). One surprising finding was that we found lower accuracy in multiple feedback condition on the hit rate compared to the single feedback or the no-feedback condition. Our findings could be interpreted as revealing negative effects of the multiple feedback condition on the hit rate compared to the single feedback or the no-feedback control condition; however, due to the inconsistent pattern of results, this interpretation should be avoided.

Our results are not in line with the results of training programs that have shown positive effects of training on assessment accuracy even though some of those programs used a similar feedback intervention to that used in our study (Baird et al., 2017; Chamberlain & Taylor, 2011; Choi & Wolfe, 2020; Raczynski et al., 2015; Rethinasamy, 2021). However, those training programs varied in other aspects of the training and did not compare the feedback group to a control group, which may explain the difference in the results. Further, those studies combined feedback with other interventions that elaborated on the feedback, such as discussions with other training participants (e.g., Choi & Wolfe, 2020) or expert talks (e.g., Baird et al., 2017). Maybe more explanations of the expert judgments are needed to change teachers' assessment standards and to improve assessment accuracy; this fits with the finding in student samples that evaluative feedback, such as simple scores or grades, has been shown to have smaller effects on performance and motivation than more elaborated written comments (Koenka et al., 2019).

Regarding assessment motivation, our results showed positive effects for the single and multiple feedback groups, in line with our hypotheses. Participants who compared their assessments with the expert judgments once or multiple times were more interested in assessing texts than participants who did not. This result aligns with studies that have shown positive effects of diagnostic training on confidence in assessment (DeLuca et al., 2013; Dempsey et al., 2004). The feedback in our study may have fostered assessment motivation by reducing the (perceived) assessment difficulty by showing teachers that their assessments were mostly close to the expert judgments. Feedback could thus inspire teachers to participate in further interventions such as discussions about assessments and, thus, could make assessment training more effective. This finding adds to the previously small body of literature on promoting assessment motivation and shows that information about one's own assessments is of interest even to experienced teachers and can help them to become more motivated to assess accurately.

Limitations

When interpreting the findings of our study, five limitations should be given special consideration. First, the idea of our study was to help teachers assess student performance in a way that is conducive to student learning, for example, by generating effective feedback based on accurate assessments. In the literature, assessments that promote learning are usually operationalized using expert judgments (e.g., Keller et al., 2019), which is far from being perfect (see Hennes et al., 2022, for a detailed discussion). We followed this approach but would like to point out that it is unclear whether expert judgments are a suitable operationalization of assessments that are conducive to student learning. To the best of our knowledge, no empirical study has yet investigated which assessments of student performance are conducive to learning (see Anders et al., 2010; Förster et al., 2022; Stang & Urhahne, 2016, for evidence on judgments of students' characteristics). However, in the context of feedback research, which is based on the idea that feedback should close the gap between actual performance and a target performance (Black & William, 2009), it can be assumed, from the perspective of educational theory, that a correctly identified gap is a prerequisite for the effectiveness of feedback.

Second, we did not assess baseline situational interest, which is why we cannot be sure that the differences between the groups were caused by our intervention. We assume that situational interest did not differ between the groups before the intervention because we used a randomized design to allocate conditions and the groups did not differ on any of the variables measured (see ESM 1, Table S1). However, we have no empirical evidence to support this claim for situational interest.

Third, our results on the accuracy feedback on teachers' text assessments cannot be transferred to the literature on teachers' judgments of student characteristics because the cognitive process differs between the two situations (see Herppich et al., 2018 and Loibl et al., 2020, for a detailed discussion). When assessing texts, teachers must perform a cognitively demanding activity in which they process new and complex information. In contrast, retrieving their knowledge about student characteristics to assess how the student will perform on an upcoming achievement test (see Südkamp et al., 2012) is less cognitively demanding. It can be assumed that this difference in cognitive load affects the effectiveness of feedback, which limits the transferability of the findings. This means that our findings apply only to the context of teacher assessments and that conclusions about the judgment of student characteristics in the tradition of the judgment accuracy literature cannot be drawn (Südkamp et al., 2012).

Fourth, we performed power analyses, assuming effect sizes of $d = 0.40$, but the differences between the groups were smaller (about $d = 0.15$). Accordingly, the power of our study might have been too low to show effects on assessment accuracy between groups and this reduces the chance that our statistically significant result reflects a true effect (Ioannidis, 2005). We assumed the effect size of $d = 0.40$ in line with a meta-analysis that included an intervention on writing assessment and other domains (Chernikova et al., 2019). Perhaps the assessment of written performance is particularly difficult to learn because of its complexity, which is why our intervention showed smaller effects.

Fifth, regarding our variation, we decided not to give any feedback in the control group, so we do not know whether the effect we found was due to the fact that the chosen feedback was particularly interesting or was merely due to the fact that something other than the text was presented. We chose this control group because, in school practice, teachers do not receive any feedback. Another variation issue is that we varied the presentation of the feedback from one time to five times between the two intervention groups. In the training programs cited above (e.g., Choi & Wolfe, 2020), participants received the feedback more than 50 times. Thus, it is possible that, with more frequent presentations in the training sessions, the feedback would be effective.

Conclusion

Student teachers' and experienced teachers' accuracy in assessing students' writing performance needs to be improved (e.g., Birkel & Birkel; Jansen et al., 2021b) but established assessment training programs are time-consuming and might not fit teachers' schedules. Using fast and easy-to-implement assessment training, this study investigated the effects of accuracy feedback as a training component. Our feedback provided teachers with a comparison between their assessments and correct assessments, operationalized by expert judgments. Teachers showed interest in the feedback but did not change their assessment behavior in response. Our interpretation is that teachers saw the feedback more as an affirmation of their competence, which they found interesting, rather than as an indication of inaccurate assessments that necessitates a change in their assessment practices. In sum, one message of our study is that our feedback could be used to help teachers to stay engaged during long periods of assessment tasks, for example, in online learning settings.

Electronic supplementary material

The electronic supplementary material (ESM) is available with the online version of the article at https://doi.org/10.1024/1010-0652/a000365
ESM 1. Figures S1–S6; Tables S1–S3 (PDF)

References

Anders, Y., Kunter, M., Brunner, M., Krauss, S., & Baumert, J. (2010). Diagnostische Fähigkeiten von Mathematik Lehrkräften und ihre Auswirkungen auf die Leistungen ihrer Schülerinnen und Schüler [Mathematics teachers' diagnostic skills and their impact on students' achievements]. *Psychologie in Erziehung und Unterricht, 57*(3), 175 – 193. https://doi.org/10.2378/peu2010.art13d

Baird, J.-A., Meadows, M., Leckie, G., & Caro, D. (2017). Rater accuracy and training group effects in Expert-and Supervisor-based monitoring systems. *Assessment in Education, 24*(1), 44 – 59. https://doi.org/10.1080/0969594X.2015.1108283

Baumert, J., & Kunter, M. (2006). Stichwort: Professionelle Kompetenz von Lehrkräften [Keyword: Teachers' professional competencies]. *Zeitschrift Für Erziehungswissenschaft, 9*(4), 469 – 520. https://doi.org/10.1007/s11618-006-0165-2

Biber, D., Nekrasova, T., & Horn, B. (2011). The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis. *ETS Research Report Series, 2011*(1), i-99.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (Formerly: Journal of Personnel Evaluation in Education), 21*(1), 5 – 31. https://doi.org/10.1007/s11092-008-9068-5

Canz, T. (2015). *Validitätsaspekte bei der Messung von Schreibkompetenzen*. [Validity aspects in the measurement of writing competencies] https://doi.org/10.18452/17333

Canz, T., Hoffmann, L., & Kania, R. (2020). Presentation-mode effects in large-scale writing assessments. *Assessing Writing*.

Chamberlain, S., & Taylor, R. (2011). Online or face-to-face? An experimental study of examiner training. *British Journal of Educational Technology, 42*(4), 665 – 675. https://doi.org/10.1111/j.1467-8535.2010.01062.x

Chernikova, O., Heitzmann, N., Fink, C., Timothy, V., Seidel, T., & Fischer, F. (2019). Facilitating diagnostic competences in higher education-a meta-analysis in medical and teacher education, 1 – 40. https://doi.org/10.1007/s10648-019-09492-2

Chernikova, O., Heitzmann, N., Opitz, A., Seidel, T., & Fischer, F. (2022). A theoretical framework for fostering diagnostic competences with simulations in higher education. In F. Fischer & A. Opitz (Eds.), *Learning to Diagnose with Simulations: Examples from Teacher Education and Medical Education* (pp. 5 – 16). Springer, Cham.

Chernikova, O., Heitzmann, N., Stadler, M., Holzberger, D., Seidel, T., & Fischer, F. (2020). Simulation-based learning in higher education: a meta-analysis. *Review of Educational Research*, 0034654320933544. https://doi.org/10.3102/0034654320933544

Choi, I., & Wolfe, E. W. (2020). The impact of operational scoring experience and additional mentored training on raters' essay scoring accuracy. *Applied Measurement in Education, 33*(3), 210 – 222. https://doi.org/10.1080/08957347.2020.1750404

Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool. *Discourse Processes, 51*(5 – 6), 511 – 534. https://doi.org/10.1080/0163853X.2014.910723

DeLuca, C., Chavez, T., & Cao, C. (2013). Establishing a foundation for valid teacher judgement on student learning: The role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice, 20*(1), 107 – 126. https://doi.org/10.1080/0969594X.2012.668870

Dempsey, M. S., PytlikZillig, L. M., & Bruning, R. H. (2009). Helping preservice teachers learn to assess writing: Practice and feedback in a Web-based environment. *Assessing Writing, 14*(1), 38 – 61. https://doi.org/10.1016/j.asw.2008.12.003

Elliott, J., Lee, S. W., & Tollefson, N. (2001). A reliability and validity study of the Dynamic Indicators of Basic Early Literacy Skills – Modified. *School Psychology Review, 30*(1), 33 – 49.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175 – 191. https://doi.org/10.3758/BF03193146

Fleckenstein, J., Meyer, J., Jansen, T., Keller, S., & Köller, O. (2020). Is a Long Essay Always a Good Essay? The Effect of Text Length on Writing Assessment. *Frontiers in Psychology, 11*, 2493. https://doi.org/10.3389/fpsyg.2020.562462

Förster, N., Humberg, S., Hebbecker, K., Back, M. D., & Souvignier, E. (2022). Should teachers be accurate or (overly) positive? A competitive test of teacher judgment effects on students' reading progress. *Learning and Instruction, 77*, 101519. https://doi.org/10.1016/j.learninstruc.2021.101519

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81 – 112. https://doi.org/10.3102/003465430298487

Heitzmann, N., Seidel, T., Opitz, A., Hetmanek, A., Wecker, C., Fischer, M., Ufer, S., Schmidmaier, R., Neuhaus, B., Siebeck Matthias, Stürmer, K., Obersteiner, A., Reiss, K., Girwidz, R., & Fischer, F. (2019). Facilitating diagnostic competences in simulations: A conceptual framework and a research agenda for medical and teacher education. *Frontline Learning Research, 7*(4), 1 – 24. https://doi.org/10.14786/flr.v7i4.384

Hennes, A.-K., Schmidt, B. M., Yanagida, T., Osipov, I., Rietz Christian, & Schabmann Alfred (2022). Meeting the Challenge of Assessing (Students') Text Quality: Are There any Experts Teachers Can Learn from or Do We Face a More Fundamental Problem? *Psychological Test and Assessment Modeling, 64*, 272 – 303.

Herppich, S., Praetorius, A.-K., Förster, N., Karst, K., Leutner, D., Behrmann, L., Böhmer, M., Ufer, S., & Südkamp, A. (2018). Teachers' assessment competence: Integrating knowledge-, process-, and product-oriented approaches into a competence-oriented conceptual model. *Teaching and Teacher Education*(76), 1 – 13. https://doi.org/10.1016/j.tate.2017.12.001

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

Jansen, T. & Möller, J. (2022). Teacher Judgments in School Exams: Influences of Students' Lower-Order-Thinking Skills on the Assessment of Students' Higher-Order-Thinking Skills. *Teacher and Teaching Education*. 103616. https://doi.org/10.1016/j.tate.2021.103616

Jansen, T., Vögelin C., Machts, N., Keller, S. & Möller, J. (2021a). Don't Just Judge the Spelling! The Influence of Spelling on Assessing Second Language Student Essays. *Frontline Learning Research*. https://doi.org/10.14786/flr.v9i1.541

Jansen, T., Vögelin, C., Machts, N., Keller, S., Köller, O., & Möller, J. (2021b). Judgment accuracy in experienced versus student teachers: Assessing essays in English as a foreign language. *Teaching and Teacher Education, 97*, 103216. https://doi.org/10.1016/j.tate.2020.103216

Jansen, T., Vögelin, C., Machts, N., Keller, S., & Möller, J. (2019). Das Schülerinventar ASSET zur Beurteilung von Schülerarbeiten im Fach Englisch: Drei experimentelle Studien zu Effekten der Textqualität und der Schülernamen [The ASSET student inventory for assessing student texts in English: three experimental studies on effects of text quality and student names.]. *Psychologie in Erziehung und Unterricht, 66*(4), 303 – 315. https://doi.org/10.2378/peu2019.art21d

Keller, S., Fleckenstein, J., Krüger, M., Rupp, A. A., & Köller, O. (2020). English Writing Skills of Students in Upper Secondary Education: Results from an Empirical Study in Switzerland and Germany. *Journal of Second Language Writing*. Advance online publication. https://doi.org/10.1016/j.jslw.2019.100700

Klug, J., Bruder, S., & Schmitz, B. (2016). Which variables predict teachers diagnostic competence when diagnosing students' learning behavior at different stages of a teacher's career? *Teachers and Teaching, 22*(4), 461 – 484.

Koenka, A. C., Linnenbrink-Garcia, L., Moshontz, H., Atkinson, K. M., Sanchez, C. E., & Cooper, H. (2019). A meta-analysis on the impact of grades and comments on academic motivation and achievement: a case for written feedback. *Educational Psychology*, 1 – 22. https://doi.org/10.1080/01443410.2019.1659939

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine, 15*(2), 155 – 163. https://doi.org/10.1016/j.jcm.2016.02.012

Kron, S., Sommerhoff, D., Achtner, M., Stürmer, K., Wecker, C., Siebeck, M., & Ufer, S. (2022). Cognitive and Motivational Person Characteristics as Predictors of Diagnostic Performance: Combined Effects on Pre-Service Teachers' Diagnostic Task Selection and Accuracy. *Journal Für Mathematik-Didaktik, 43*(1), 135 – 172. https://doi.org/10.1007/s13138-022-00200-2

Leckie, G., & Baird, J.-A. (2011). Rater effects on essay scoring: A multilevel analysis of severity drift, central tendency, and rater experience. *Journal of Educational Measurement, 48*(4), 399 – 418. https://doi.org/10.1111/j.1745-3984.2011.00152.x

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

Merk, S., Ophoff, J. G., & Kelava, A. (2023). Rich data, poor information? Teachers' perceptions of mean differences in graphical feedback from statewide tests. *Learning and Instruction, 84*, 101717. https://doi.org/10.1016/j.learninstruc.2022.101717

McElvany, N., & Artelt, C. (2009). Systematic reading training in the family: Development, implementation, and initial evaluation of the Berlin Parent–Child Reading Program. *Learning and Instruction, 19*(1), 79 – 95. https://doi.org/10.1016/j.learninstruc.2008.02.002

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

Möller, J., Zitzmann, S., Helm, F., Machts, N., & Wolff, F. (2020). A meta-analysis of relations between achievement and self-concept. *Review of Educational Research, 90*(3), 376 – 419. https://doi.org/10.3102/0034654320919354

NAEP. *National Assessment Governing Board (2011). Writing Framework for the 2011 National Assessment of Educational Progress. U.S.* Department of Education, Washington, D.C.

NAEP. (2011). *National Assessment Governing Board (2011). 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.

NAEP. (2012). *National Center for Education Statistics (2012). The Nation's Report Card: Writing 2011. Institute of Education Sciences, U.S.* U.S. Department of Education.

Narciss, S. (2007). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, & J. J. G. van Merrienboer & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (pp. 125 – 144). Lawrence Erlbaum Associates.

Narciss, S. (2018). Feedbackstrategien für interaktive Lernaufgaben. [Feedback strategies for interactive learning tasks] In S. Kracht, A. Niedostadek, & P. Sensburg (Eds.), *Springer Reference Psychologie. Praxishandbuch Professionelle Mediation* (pp. 1 – 24). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-54373-3_35-1

Raczynski, K. R., Cohen, A. S., Engelhard Jr, G., & Lu, Z. (2015). Comparing the effectiveness of self-paced and collaborative frame-of-reference training on rater accuracy in a large-scale writing assessment. *Journal of Educational Measurement, 52*(3), 301 – 318. https://doi.org/10.1111/jedm.12079

Rethinasamy, S. (2021). The Effects of Different Rater Training Procedures on ESL Essay Raters' Rating Accuracy. *SOCIAL SCIENCES & HUMANITIES*, 401 – 419. https://doi.org/10.47836/pjssh.29.S3.21

Rezaei, A. R., & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18 – 39. https://doi.org/10.1016/j.asw.2010.01.003

Rotgans, J. I., & Schmidt, H. G. (2014). Situational interest and learning: Thirst for knowledge. *Learning and Instruction, 32*, 37 – 50. https://doi.org/10.1016/j.learninstruc.2014.01.002

Royal-Dawson, L., & Baird, J.-A. (2009). Is teaching experience necessary for reliable scoring of extended English questions? *Educational Measurement: Issues and Practice, 28*(2), 2 – 8. https://doi.org/10.1111/j.1745-3992.2009.00142.x

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119 – 144. https://doi.org/10.1007/BF00117714

Scannell, D. P., & Marshall, J. C. (1966). The effect of selected composition errors on grades assigned to essay examinations. *American Educational Research Journal, 3*(2), 125 – 130.

Sommerhoff, D., Codreanu, E., Nickl, M., Ufer, S. & Seidel, T. (2022) Pre-service teachers' learning of diagnostic skills in a video-based simulation: Effects of conceptual vs. interconnecting prompts on judgment accuracy and the diagnostic process. *Learning and Instruction.* 10.1016/j.learninstruc.2022.101689

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153 – 189. https://doi.org/10.3102/0034654307313795

Skar, G. B., & Jølle, L. J. (2017). Teachers as raters: Investigation of a long-term writing assessment program. *L1 Educational Studies in Language and Literature, 17, Open Issue*(Open Issue), 1 – 30. https://doi.org/10.17239/L1ESLL-2017.17.01.06

Spinath, B. (2005). Akkuratheit der Einschätzung von Schülermerkmalen durch Lehrer und das Konstrukt der diagnostischen Kompetenz: Accuracy of teacher judgments on student characteristics and the construct of diagnostic competence. [Accuracy of Teacher Judgments on Student Characteristics and the Construct of Diagnostic Competence]. *Zeitschrift für Pädagogische Psychologie, 19*(1/2), 85 – 95. https://doi.org/10.1024/1010-0652.19.12.85

Spinath, B., & Steinmayr, R. (2012). The roles of competence beliefs and goal orientations for change in intrinsic motivation. *Journal of Educational Psychology, 104*(4), 1135 – 1148. https://doi.org/10.1037/a0028115

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743 – 762. https://doi.org/10.1037/a0027627

Südkamp, A., Möller, J., & Pohlmann, B. (2008). Der Simulierte Klassenraum: Eine experimentelle Untersuchung zur diagnostischen Kompetenz.: [The Simulated Classroom: An Experimental Study on Diagnostic Competence]. *Zeitschrift für Pädagogische Psychologie, 22*(34), 261 – 276. https://doi.org/10.1024/1010-0652.22.34.261

Stang, J., & Urhahne, D. (2016). Stabilität, Bezugsnormorientierung und Auswirkungen der Urteilsgenauigkeit [Stability, reference norm orientation, and effects of judgment accuracy]. *Zeitschrift für Pädagogische Psychologie, 30*, pp. 251 – 262 https://doi.org/10.1024/1010-0652/a000190

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

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*, 1 – 17. https://doi.org/10.1080/09571736.2018.1522662

Vögelin, C., Jansen, T., Keller, S., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgments 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., Machts, N., & Möller, J. (2020). Organisational quality of ESL argumentative essays and its influence on pre-service teachers' judgments. *Cogent Education, 7*(1), 1760188. https://doi.org/10.1080/2331186X.2020.1760188

Vu, T., Magis-Weinberg, L., Jansen, B. R. J., van Atteveldt, N., Janssen, T. W. P., Lee, N. C., van der Maas, H. L. J., Raijmakers, M. E. J., Sachisthal, M. S. M., & Meeter, M. (2021). Motivation-Achievement Cycles in Learning: A Literature Review and Research Agenda. *Educational Psychology Review*, 1 – 33. https://doi.org/10.1007/s10648-021-09616-7

History
Received April 10, 2022
Accepted after revision May 4, 2023
Published online June 12, 2023

Conflict of Interest
We have no conflicts of interest to disclose.

Funding
The research in this manuscript was funded by the German Research Foundation [Grant Nr. Mo648 / 25-1].
Open access publication enabled by IPN – Leibniz Institute for Science and Mathematics Education, Kiel, Germany.

ORCID
Thorben Jansen
https://orcid.org/0000-0001-9714-6505
Jennifer Meyer
https://orcid.org/0000-0002-5714-3198
Jens Möller
https://orcid.org/0000-0003-1767-5859

Thorben Jansen
University of Kiel
Leibniz Institute for Science and Mathematics Education
Kuhnkestraße 2
24118 Kiel
Germany
tjansen@ipn.uni-kiel.de