Synthesis writing

Teaching high school students how to read, plan, draft, and revise

van Ockenburg, L.

Publication date
2022

CHAPTER 6

STUDENTS' ASSESSMENT OF SYNTHESIS TEXT QUALITY

The Effect of a One-Hour Assessment Training Session[*]

Abstract. Usually, teachers assess their students' writing assignments themselves, but perhaps students could also play a role in (co-)assessing their peers' text quality. To this end, we examined the extent to which we can teach students to assess the quality of synthesis texts in the same way that experienced raters (experts) do. Therefore, we examined the degree of agreement between student and expert assessments in four areas: internal consistency, strictness, external coherence, and meaning. The results showed that student and expert agreement (internal consistency) did not differ at any measurement time and for any condition. Similarly, the conditions did not differ with respect to agreement on quality ranking (external coherence). Participation in an assessment training for students had no effect on these two items. However, students did give the texts a significantly higher holistic rating than experts. After training, students' assessments were still higher, but the difference with experts' assessments had become significantly smaller compared to the pretest and disappeared completely when assessing texts on a different topic. In addition, the meaning of the holistic student assessment, i.e., the weighting of various aspects of text quality in the holistic final assessment, did not (completely) correspond to the holistic expert assessment at all measurement points, but after training it started to correspond more closely to the experts' weighting. Having students (co-)assess their peers' texts therefore seems to be a realistic option.
Keywords: writing education, synthesis texts, text quality, assessment, interrater agreement

1 INTRODUCTION

When students comment on each other's texts, this not only benefits the teacher, but also the students themselves. There is even evidence that students learn more from commenting on each other's texts than from receiving feedback (Chanski & Ellis, 2017, p. 58). The Dutch foundation for curriculum development [Stichting Leerplanontwikkeling, SLO], also advises, in its handbook for Dutch language and literature school exams havo/vwo (Bonset et al., 2012, p. 48), to let students comment on each other's rough drafts and thus learn more about the aspects on which text quality is based. A logical next step would be to subsequently involve students in the assessment of the final product as well. This step, summative student assessments of text quality, is the focus of the current article.

---

[*] *This chapter is a somewhat adapted version of Van Ockenburg, L., Van Weijen, D., & Rijlaarsdam, G. (2022). Beoordeling van schrijfopdrachten door leerlingen. Effecten van een korte beoordelaarstraining. [Students' assessment of synthesis text quality. The Effect of a One-Hour Assessment Training Session.] Under review.*

1.1 *Students as text raters*

Since 2016, we have been conducting research on ways to teach ninth grade secondary school students to write synthesis texts. Synthesis texts are based on sources and are a representative and well-integrated representation of the sources that can be easily understood by a reader who has not read the sources (Klein & Boscolo, 2016). We developed and tested a learning unit on synthesis texts at three secondary schools in the Netherlands (see Chapter 5). To determine the effect of the learning unit on students' text quality, 396 synthesis texts had to be assessed. Because synthesis texts are relatively unknown in secondary education, we developed an assessment tool. A justification for the development of this assessment instrument can be found in Chapter 3. The instrument consisted of three components:

1. An animated video explaining what synthesis texts are,
2. An assessment tool that described the five quality levels of four aspects: information, integration, structure, and style, and,
3. A text scale with annotated benchmark texts for holistic assessment (on a scale of 0-100).

We instructed raters to first watch the animation video, then study the student assignment and associated resources, then carefully review the text scale to get an idea in advance of the possible variation in text quality. The scale consisted of four benchmark texts of increasing holistic quality. Each benchmark text was accompanied by an explanation of the quality for each aspect. Finally, we asked the reviewers to first rate each text on the four aspects and then assign a holistic score.

Each text was independently assessed by a group of three raters from a panel of 25 experts (Dutch language teachers and writing researchers). We determined the inter-rater agreement according to the method developed by Van den Bergh and Eiting (1989). This turned out to be sufficient for research purposes ($\rho = .70$; see Chapter 5). Thus, the assessment tool was sufficiently useful when applied by experts.

Since students assess (parts of) each other's synthesis texts in the learning unit, we were curious whether students can actually do this adequately. Therefore, we set up an experiment to investigate to what extent students are able to adequately assess each other's texts using the expert assessment tool and whether a short training could improve students' assessment skills. If results suggest that students are able to rate texts in a way that corresponds to the rating

experts would give, then student assessment can no longer just be used as a proven learning activity that enables students to acquire knowledge about text quality, but also as a means for summative assessment.

In addition, the results of this study may shed more light on an issue surrounding the assessment of writing skills, which was introduced in the Netherlands by Wesdorp (1974, 1981), and further pursued by Schoonen (2012) and Van den Bergh et al. (2012). In short, they proposed that writing skills can only be reliably assessed if several texts are assessed per student and if each assessment is conducted by multiple, independent raters.

In a school setting, however, assessment by multiple teachers is unfeasible: language departments rarely can have all the texts written by all their students assessed by three teachers. Expanding the number of writing assignments per student per term is also not feasible without increasing the teacher's work load. But what if we could involve students in the assessment procedure?

We are aware of one experiment in the Netherlands in which students were involved in summative assessment: assessing goal- and audience-oriented documented writing for a school examination grade. We use the data reported by this experiment as a frame of reference. In two articles, Rijlaarsdam and Blok (Rijlaarsdam and Blok, 1981; Blok and Rijlaarsdam, 1981) reported on an experiment in which students' texts for a school examination was assessed by a panel consisting of the teacher and two students. See Rijlaarsdam and Blok (1981) for the didactic background. The results of the experiment showed that a jury consisting of a teacher and two students is reasonably reliable (reliability of .62 for holistic text quality), and certainly more reliable than the assessment by a single teacher. In addition, Blok and Rijlaarsdam also analyzed the validity of the assessments: do the raters understand the aspects to be assessed in the same way?

The purpose of this study was to determine the extent to which student and expert assessments agree, in terms of reliability and validity, and the extent to which a brief training affects reliability and validity. To this end, we will answer the following research questions:

1. To what extent does a brief training affect agreement between student and expert assessments?
2. To what extent does the topic of synthesis texts affect the agreement between student and expert assessments?

We will examine each research question in four areas:

a) Internal consistency (within conditions): do holistic student and expert assessments of texts differ in reliability, if we compare the average mutual agreement between two students on the one hand and two experts on the other?
b) Strictness: do students and experts differ in their standards, evidenced by differences in their mean holistic quality scores?
c) External coherence (between conditions): do students and experts differ in their idea of text quality, when looking at the average agreement on ranking between experts and students?
d) Meaning: do students' and experts' holistic assessments rely to the same extent on the same aspects of text quality?

We answered the questions using student and expert scores on the same sets of texts. Each set consisted of 124 synthesis texts written by students in 10 ninth grade (3-vwo) classes from three different high schools in the Netherlands. The texts had to be about 200 words long and were based on three short complementary sources (number of words per source M = 188.9, SD = 55). The first set consisted of texts about endangered wild animals in Africa and the second set of texts about artificial food coloring (E-numbers). The assignments were based on synthesis tasks that had been developed and tested as part of a national assessment study of synthesis tasks in upper secondary schools (Vandermeulen et al., 2020).

2 METHOD

2.1 *Participants*

Seventy-five ninth and 10th grade students from the teacher-researcher's school participated in the study. Students' characteristics and distribution across conditions are provided in Table 6.1. Students volunteered to attend an information session outside of school hours after an appeal by the teacher-researcher via the school's content management system (Magister) and in class. Upon participation, they received a small financial compensation in the form of a gift voucher, but only if they had completed all research activities. All participants and their parents gave active consent for their participation.

Table 6.1 Characteristics of participants by condition

Condition	Gender		Age (years) *m* (*sd*)	Classes*
	Boy	Girl		
1	7	18	14.28 (0.5)	10
2	13	12	14.08 (0.4)	6
3	8	17	14.08 (0.6)	9
Total	28	47	14.15 (0.6)	10

* Number of different classes represented in the condition

2.2 *Research Design*

The research design was based on the *Solomon four group design*: a pretest-posttest design with a control group and random assignment to conditions. Table 6.2 shows this design which includes three measurement occasions and three conditions: Condition 1: a pretest, followed by a training; Condition 2: a pretest, but no training; Condition 3: no pretest, only a training. A complete *Solomon four group design* also includes a fourth condition, involved only in the posttest. However, such an extension was not feasible for this study because there were too few students participating in the study to create four conditions. A lower number of participants within each condition would likely be detrimental to the reliability of the analyses. In addition, the posttest was nearly identical to the pretest. The only difference was that the participants rated a different sample from the same collection of texts that we used for the pretest. Thus, no differences between pretest and posttest were expected to be due to a different form or content of the two measurements.

Table 6.2 Research Design

Condition	*n*	Pretest (topic A)	Training	Posttest (topic A)	Transfer test (topic B)
1	25	o	x	o	o
2	25	o		o	o
3	25		x	o	o

x = participation in training; o = text assessment

In this design, we used the pretest and posttest to measure the effect of training on topic A on students' assessments of texts on topic A (comparison of Conditions 1 and 2). With an extension of this design, namely the addition of a transfer test after the posttest, we measured whether agreement on topic A during the posttest, is maintained when students assess texts on another, untrained topic (topic B).

A second extension of the design was the addition of a third condition in which students participated in the training without participating in the pretest. This allowed us to determine if the pretest resulted in a possible learning effect. Without this control, we could not determine whether such a training is directly usable in educational practice. If a positive effect of the training is found, then that effect is likely to have been obtained after a pretest (unintentional practice with topic A) and a training (intended practice with topic A). A learning effect could be expected from the pretest: it is possible that students who have gained experience as raters during the pretest, benefitted more from the training because of this experience and assessed texts more effectively during the posttest than students who followed the training without this experience. Thus, Condition 3 enabled us to measure the "pure" effect of the training.

2.3 *Procedure*

Prior to their participation in the study, all interested students were briefed during a live meeting at school in small groups. Here, the first author, a Dutch teacher at the school in question, went through the research procedure with them, i.e., which tasks they would have to carry out within which time frame. After this meeting, students could decide whether they definitely wanted to participate or not.

Subsequently, four research activities took place in four consecutive weeks: three assessment sessions and one training session. Each activity lasted approximately 60 minutes and was scheduled flexibly. The three assessment sessions (pre-, post-, and transfer tests) were conducted by the student raters at home. At the beginning of the week, the students received 15 texts on paper that they had to assess. Once they finished assessing them, they could submit their assessments digitally. The students in Condition 3, the condition without a pretest, received an alternative task to perform instead of the pretest which consisted of a literary reading comprehension task. The task required a similar time investment as the assessment task the other students had to perform, but was not related to the study in terms of content. Two more rounds of assessment followed the pretest. During the posttest, the students each assessed a different set

of texts from the total set we drew from for the pretest. Finally, for the transfer test, the students assessed a set of synthesis texts on a different topic.

The training sessions for both experimental Conditions 1 and 3 took place at school, between the pre- and posttests. The training was offered at six different times and students signed up for a time that fitted well into their school timetable. Two students who were in quarantine during the week in question due to Covid-19, followed the training online via Microsoft Teams. The students from control Condition 2 (pretest, no training) also enrolled and were present at school during the training session, but spent that hour in the school library doing the same task as the students from Condition 3 had done before: A literary reading comprehension task.

2.4 *Content of assessment training*

The content of the assessment training was based on an assessment training for teachers (Echten et al., 2020) that we developed as part of the Dutch Writing for the School Exam project (https://didactieknederlands.nl/op-naar-een-beter-schoolexamen-schrijfvaardigheid/) in collaboration with the Werkgroep Onderzoek en Didactiek Nederlands [Working Group Research and Education of Dutch, WODN] and The Dutch foundation for curriculum development [Stichting Leerplanontwikkeling, SLO]. The aim of the original training was to allow teachers to discuss their views on text quality by comparing and ranking texts of different quality, thus creating a (more) shared view on text quality, which would in turn result in more reliable text quality assessments. The reliability of the assessments can be increased by using assessment criteria and scales with benchmark texts. The training was based on these two pillars. First, we wanted to train raters to use criteria that would give them tools to assess every written product in the same way. An analytical rating scheme, for example, can define criteria that are taken into account in the assessment (Coertjens et al., 2017). Second, a benchmark text scale can also facilitate text quality assessment (Koster et al., 2018; Wesdorp, 1981). Such a scale consists of a series of benchmark texts of increasing quality, each with a substantiated assessment. A text scale can prevent certain assessment problems, such as the sequence effect (which means that a good text that follows after a number of weaker texts is rated higher than if it would follow after another good text and v.v.), norm-shifting (which means that a rater (unconsciously) adjusts his assessment to the level of the texts), or a signifier effect (different raters value different text aspects and therefore their ratings of the same text differ) (Koster et al. 2018; Pollmann et al., 2012; Wesdorp, 1981).

The aim of this study was to make students' assessments more consistent with experts' assessments. With this goal in mind, the content of the training was modified and more guiding than the original teacher training. Table 6.3 summarizes the content of the student training. In several rounds of assessment, students compare their own assessments of sample texts with those of experts, and these experts' assessments are explained. At the end of the training, it should be clear how experts assess text quality and why they do so, so that students can adjust their own assessments accordingly.

Table 6.3 Structure and overall content of the student assessment training

Round	Min.	How?*	Contents
1	10	I	Students individually place three informational sample texts in order of quality and provide arguments for this ranking.
2	10	G	Students compare their own rankings with expert rankings and discuss how the aspects of information, integration, structure, and style factor into assessing text quality.
3	5	I	Students individually add two new sample texts to the expert ranking from round 2.
4	5	G	Students compare their own and the experts' rankings, discuss any differences, and determine whether the quality aspects from Round 2 are clear enough or need further clarification.
5	10	I	Students determine the score for each aspect on a scale of 1 to 5 (analytical) and the holistic score (1-100) for the five sample texts from rounds 1-4.
6	10	G	Students compare their own analytical and holistic assessments with the experts' assessments and discuss any differences.

* I = Individual; G = Group discussion led by trainer

3 RESULTS

3.1 *RQ 1: Agreement after training*

The first question we wanted to answer was related to the extent to which participation in a brief training affected agreement between expert and student

assessments in terms of internal consistency, strictness, external consistency, and significance.

(a) Internal consistency

Table 6.4 shows the reliability of the holistic assessments within the three groups that generated scores for the pretest: Experts and Conditions 1 and 2. The correlations between two raters from the same condition ranged from .33 (experts) to .26 (Condition 2 students). There were no significant differences in reliability at any of the three measurement occasions (test statistic $z = .25$, $p = .40$; Lenhard & Lenhard, 2014), i.e., the correlation between two experts and two students did not differ significantly. Thus, the training had no effect on the assessments' internal consistency.

Table 6.4 Internal consistency: Generalization coefficient of the correlation between two raters from one condition for holistic text quality

			Measurement occasions		
			Pretest (M1)	Posttest (M2)	Transfer test (M3)
Topic			A	A	B
Condition	Pretest	Training			
Experts			.33	.33	.33
Students Condition 1	yes	yes	.33	.32	.23
Students Condition 2	yes	no	.26	.35	.27
Students Condition 3	no	yes	-	.37	.30

(b) Strictness

Table 6.5 shows the mean text quality assessments per condition per measurement occasion. A multilevel analysis of the holistic assessments for pretest (three conditions: experts and two student conditions), showed that the height of the mean assessments between experts and students differed significantly ($F(2, 991{,}083) = 30.72$, $p < .001$). Pairwise comparisons showed that students' assessments were significantly higher than experts' assessments (both comparisons $p < .001$), while the two student conditions did not differ ($p = .89$).

Table 6.5 Strictness: Mean (and standard deviation) of holistic assessments (scale 0-100) for Experts and for students' conditions

			Measurement occasions		
			Pretest (M1)	Posttest (M2)	Transfer test (M3)
Topic			A	A	B
Condition	Pretest	Training			
Experts			52.9 (13.2)	52.9 (13.2)	66.4 (10.9)
Ss' Condition 1	yes	yes	60.3 (12.4)	57.8 (11.7)	67.9 (8.9)
Ss' Condition 2	yes	no	60.5 (12.7)	63.1 (12.0)	69.8 (9.9)
Ss' Condition 3	no	yes	-	57.5 (13.4)	66.8 (10.5)

A multilevel analysis of the holistic assessments at posttest (see Table 6.5, M2) showed that for this task students also assessed text quality less strictly than the experts (F(3, 1363,028)= 32.49, p < .001). However, the difference between the assessments of experts and trained students in Condition 1 for the posttest was found to be significantly smaller compared to the pretest (B = 5.17, SE = 2.25, p = .02). We were unable to check this for Condition 3, as this condition did not participate in the pretest, but at the posttest, no difference between Condition 1 (m = 57.8) and Condition 3 (m = 57.5) scores were found (p = .93). Condition 2 students, without training, assessed text quality significantly higher at the post-test than their trained peers in Conditions 1 and 3 (B = 5.61, SE =1 .59, p < .001). Thus, after training, the level of strictness of holistic student assessments corresponded more closely to that of expert assessments. In other words, the trained students had become somewhat stricter in their assessments.

(c) External coherence

Although students seem to agree with each other as much as experts do when they have to rank texts for quality (see Table 6.4), it is still possible that students and experts do not rank these texts in the same way. This can be determined on the basis of the between-group correlation. Table 6.6 shows the correlation between students and experts and provides insight into the extent to which the two groups agree in their assessments about which texts are stronger and which are weaker, without looking at the norms.

The correlation between the ranking of texts by the two groups of students who participated in the pretest (pretest .26 vs. .29; test statistic z = .11, p = .46)

did not differ significantly from each other, nor from the correlation between two experts (pretest .33 vs. .26; test statistic z = .25, p = .40): before training students' assessments about the ranking of text quality were thus as strongly related to experts' assessments as experts' assessments were to each other. Students' participation in the training did not appear to affect that correlation.

Table 6.6 External coherence: Generalization coefficient between any expert and any student

			Measurement occasions		
			Pretest (M1)	Posttest (M2)	Transfer test (M3)
Topic			A	A	B
Condition	Pretest	Training			
Experts			.33	.33	.33
Students Condition 1	yes	yes	.29	.30	.26
Students Condition 2	yes	no	.26	.28	.24
Students Condition 3	no	yes	-	.28	.27

(d) Meaning

Table 6.7 shows differences in the meaning that experts and students attributed to holistic assessments. The differences between the two student conditions were not significant at the pretest: all four aspects weighed the same (Condition 1: r = .60 vs r = .63, p = .26; Condition 2: r= .62 vs r = .67, p = .12). But the differences between students and experts were twofold. First, for experts, aspect scores for information, integration, and structure weighed more heavily than aspect assessments for style (r = .61 vs. r = .74, p = .001). Students, on the other hand, involved all four aspects equally in their assessments, but their holistic assessments correlated less strongly with the aspect scores for information, integration, and structure than for the experts (the smallest difference r = .65 vs r = .74 is significant: p = .009). Thus, the correlations for those three aspect scores with holistic assessment were significantly stronger among experts prior to the students' training session.

At the posttest, a significant difference could be observed between the two training conditions (Conditions 1 and 3) in the contribution of aspect scores to the holistic assessment, when we took the contribution in the expert panel as the criterion. For the aspects information, integration and structure, the contribution in Condition 1 is significantly smaller than the criterion (the smallest difference, r

= .75 vs r = .67 is statistically significant p = .014). For Condition 3, this is true only for the aspect of integration (p = .007). Apparently, the pretest, in which Condition 1 students participated and Condition 3 students did not, influenced the extent to which the training has an effect in this respect. Condition 2 students, who only participated in the pretest but did not receive training, scored lower on the aspects of integration and structure than the criterion (the difference for information is not significant, p = .06). For the style aspect, the contribution to the holistic assessment did not differ significantly from the criterion in all three conditions.

Table 6.7 Coherence: Correlations of aspect scores with holistic text quality

		Experts	Students 'Conditions		
			1	2	3
Pretest	Information	.77	.62	.67	
	Integration	.75	.60	.62	
	Structure	.74	.63	.65	
	Style	.61	.63	.62	
Posttest	Information	.77	.64	.72	.71
	Integration	.75	.67	.60	.66
	Structure	.74	.61	.63	.70
	Style	.61	.58	.63	.63
Transfer test	Information	.69	.66	.63	.72
	Integration	.77	.66	.67	.68
	Structure	.80	.69	.68	.68
	Style	.70	.58	.63	.61

1: Students in Condition 3 did not participate in the pretest.
Condition 1: Pretest and Training; Condition 2: Pretest; Condition 3: Training

3.2 *RQ 2: Transfer to texts with different topics*

Finally, we wanted to know whether the agreement between experts' and students' assessments changes when synthesis texts are assessed on a different topic. Regarding the internal consistency and external coherence of and between students' and experts' assessments, we had already found that there were no significant differences between conditions and measurement occasions when

answering research question 1. Thus, we will leave these two points aside here, and will answer research question 2 for the other two area's: strictness and meaning of assessment.

Strictness

For the transfer test, when the students assessed a synthesis text on a different topic, we found no difference between the assessments of the two training conditions (Conditions 1 and 3; B = 1.16, SE = 1.28, p = .36) and between the experts' assessments and the two training conditions (Condition 1: + 1.4 points relative to the holistic expert assessment, SE = 1.3, p = .25; Condition 3: + 0.32 points relative to the holistic expert assessment, SE = 1.3, p = .80). Students in Condition 2, the condition without training, still rated text quality significantly higher than the experts, even on the transfer test (+ 3.3 points, SE: 1.3, p = .009). Thus, the effect of training on the agreement between students' and experts' assessments appears to carry over into the assessment of texts on a different topic.

Meaning

The contribution of the integration and structure aspect scores to the holistic score was invariably larger in the experts' than in the students' assessments: the smallest difference for integration (r = .68 and r = .77, p = .005) was significant, and the smallest difference for structure (r = .80 and r = .69, p < .001) was also significant. For the information aspect, the contributions did not differ, and for style, only Condition 3 deviated with a smaller contribution compared to experts (p = .02).

3.3 *Summary of results*

In summary, students' assessments of holistic quality of synthesis texts were related to each other just as experts' assessments were related to each other in terms of their internal consistency (a), and students' assessments about the ranking of texts were also related to experts' assessments, i.e., in terms of their external coherence (c). Thus, participating in a rater training had no effect on these two items.

Students did differ from experts in terms of their strictness: at the pretest they rated texts higher in quality on average than experts did. Furthermore, the training led students to assess text quality more strictly than before training (Condition 1 and 3), but not yet as strictly as the experts did. Students in Condition 2, without training, rated the texts significantly higher than their peers in Conditions 1 and 3. On the transfer test, no difference could be shown for this issue between

Conditions 1 and 3 on the one hand and the experts on the other, while the assessments by Condition 2 students were still significantly higher.

At the pretest, the holistic student assessment was less strongly linked to the aspect scores for information, integration, and structure, than the expert assessment. At the posttest, we saw effects of training in Conditions 1 and 3, both of which weighed the four aspects in a more differentiated way in the holistic assessment. In particular, in Condition 3 the contribution of style became smaller relative to other aspects, as experts also weighed style slightly less. Nevertheless, even after training, the meaning of the students' holistic assessment did not fully correspond to the experts' holistic assessment.

4 DISCUSSION AND IMPLICATIONS FOR CLASSROOM PRACTICE

In this study, we investigated how participation in an assessment training affects students' assessment skills. We examined students' assessments in four area's, and in each case with a criterion formed by experts' assessments of the same texts: a) internal consistency, b) strictness, c) external coherence, and d) meaning.

When interpreting these findings, we must consider that these are relatively random students who opted to participate in this study voluntarily and were confronted with a text genre with which they were unfamiliar.

We compared their assessments with the assessments of a group consisting of random experts. In a teaching situation, students become familiar with their teacher's assessment method, and it can be assumed that this will result in a greater agreement between teacher and students in terms of what constitutes good and weaker texts than was possible to attain in this study.

Looking at the summary of results, both Conditions 1 and 3, show a high(er) level of agreement with expert assessments after training. This is good news for the implementation of assessment training in education, because a preliminary exercise in which students become familiar with the text genre to be assessed appears to be not necessary for effective training, since no difference could be demonstrated between condition 3, without pretest, and condition 1, with pretest.

The only issue on which full agreement was not reached between students and experts at any measurement occasion was the meaning of holistic assessment. Again, however, for this outcome, in a regular teaching situation, teachers establish criteria for text quality, often in consultation with their students, after which students practice applying these criteria by giving and receiving feedback on draft versions of their texts. In subsequent summative assessment, students

are then better able to apply these criteria in accordance with their own teacher's methods.

Based on the results of this study, it therefore seems quite possible to have students (co-)assess text quality summatively after a short training session. Students do not need to have any experience with writing or assessing the texts in question prior to the training, and the effect carries over into their assessment of texts on a different topic. By involving students in the assessment process in this simple way, written products can be assessed more reliably: assessment by one teacher is not reliable, assessment by three teachers is not feasible, but assessment by three students, or by one teacher and two or three students is feasible and more reliable. It might also be feasible to have students write more texts per term when they (co-)assess each other's texts. This will give a more versatile and thus a more complete picture of their writing skills without further increasing teachers' workload.